Disagreement and motivated reasoning

Hallsson, Bjørn Gunnar

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF MEDIA, COGNITION, AND COMMUNICATION

PhD Thesis

Bjørn Hallsson

Disagreement and motivated reasoning

Supervisor: Klemens Kappel

Submitted on: 28 February 2018

Name of department: Department of Media, Cognition, and Communication

Author: Bjørn Hallsson

Title: Disagreement and motivated reasoning

Supervisor: Klemens Kappel

Submitted on: 28 February 2018

Word count: 61,290

Table of contents

Acknowledgements

My biggest debt is owed to my supervisor, Klemens Kappel. Klemens has provided detailed and very helpful comments on almost all of the contents of this dissertation, plus on at least as many pages that never made it into the final draft. He has also been immensely supportive and ensured that I felt welcomed into the fold at the department and to philosophy as a field. In addition, Klemens deserves many thanks for his tireless efforts at ensuring that the philosophy section at the University of Copenhagen is a wonderful work environment, always with a pleasant atmosphere inside the office and enjoyable times to be had beyond it.

Several other people have read and provided helpful comments on parts of this dissertation. First and foremost, these include past and present members of the Social Epistemology Research Group at the University of Copenhagen. Apart from Klemens, they are: Josefine Pallavicini, Fernando Broncano-Berrocal, Giacomo Melis, Emil Frederik Lundbjerg Møller, Frederik Joakim Andersen, and several BA and MA students who have regularly attended our sessions. Emil, Fernando, Giacomo, and Josefine, who I at different times have shared an office with, deserve special thanks for making my working days a pleasure, even if many hours of them have been spent in always stimulating but not always work-related discussions. In addition to this group, Luca Zanetti and Michel Croce read, commented on, and discussed what eventually evolved into article 1 numerous times over the course of a weeklong workshop in Rome, while Tine Hindkjær Madsen and Judith van Ooijen did the same a year later for what evolved into article 3.

I have at various presentations of the contents of the dissertation received helpful comments from Mikkel Gerken, who deserves additional thanks for his support and interesting conversations when he has visited the department, as well as from Jennifer Lackey, Sandy Goldberg, Jesper Kallestrup, J. Adam Carter, David Christensen, Jon Matheson, Chris Kelp, Erik Olsson, Jesús Navarro, Michael Lynch, Ian James Kidd, Johan Gersel, Hanna Gunn, Casey Johnson, Nikolaj Nottelmann, Simon Barker, Martin Marchman Andersen, Xavier Landes, Thomas Raleigh, Thomas Grundmann, Brian Weatherson, Asbjørn Stieglich-Petersen, Matthias Skipper-Rasmussen, and several others to whom I apologize for forgetting to put them on this list.

Great thanks are also due my colleagues at the department, and in particular those who are or have been my fellow PhD students: Nana Cecilie Halmsted Kongsholm, Andreas Christiansen, Josefine Pallavicini, Tine Hindkjær Madsen, Katla Heðinsdóttir, Lucy Holt, Rikke Moresco Lange, and Linda Fønss, as well as MA students Frederik J. Andersen and Nanna Taszarek Holm. Nana and Andreas deserve extra special thanks for being there to make me feel welcome when I began my journey into philosophy. They did so wonderfully, and their company is, along with the others on this list, a continual source of pleasure whether it is in the office, at lunch, drinking wine at a curbside café in Rome, or at Chez André (also known as Andreas' apartment) being feasted.

Finally, I want to thank my friends from outside philosophy and my family for always being supportive, and for making the years spent writing this dissertation infinitely more enjoyable. I look forward to seeing more of them.

Coauthors

Article 3

Klemens Kappel coauthors article 3. Bjørn Hallsson is first author and Klemens Kappel co-author of the paper. Bjørn Hallsson has written the article and Klemens Kappel has contributed with substantial commentary and critical discussion.

Article 4

Andreas Christiansen coauthors article 4. Both authors contributed equally to the article. Bjørn Hallsson is primarily responsible for writing section 2, and for parts of section 3.3. Andreas Christiansen is primarily responsible for writing the introduction, section 3, and the conclusion.

Abstract

The dissertation explores the impact of motivated reasoning, a tendency for reasoning to proceed with the goal of construing evidence in ways that are supportive of a desired conclusion, for normative discussions in the epistemology of disagreement and political philosophy. The introduction provides a brief overview of the psychology of motivated reasoning and its consequences. One of these is belief polarization, a tendency for the beliefs of people who initially disagree to move toward the extremes after the persons view the same body of evidence. Another is that deliberation in groups with internal disagreement is an effective means of epistemic improvement. The introduction also provides overviews of the epistemology of disagreement and the role of disagreement for political legitimacy, and details the contributions of the four articles to these debates.

Article 1 responds to arguments for the conclusion that belief polarization is a rational phenomenon. It argues that, when disagreement is salient, the biased processing of evidence that results in belief polarization is incompatible with rationality, and the resulting polarized beliefs are neither reliably formed nor supported by the evidence, properly construed.

Article 2 discusses the epistemic significance of political disagreement. It shows that motivated reasoning about politically salient propositions implies that a political opposite's familiarity with relevant evidence and their intellectual virtues are inversely correlated with their perceived probability of being right, conditional on disagreement. This presents us with a puzzle in determining how significant such disagreements are, one that cannot be escaped by denying that political disagreements in general are epistemically significant.

Article 3 discusses the impact of the beneficial effects of collective deliberation in groups with internal disagreement for the epistemic significance of discovered disagreement. It argues that these benefits can provide one with epistemic reason to maintain confidence in the face of discovered disagreement when doing so promotes epistemically fruitful deliberation.

Article 4 discusses the impact of motivated reasoning in defense of our political or cultural values for the legitimacy of democratic decision-making. It addresses the extent to which democratic authorities should be responsive to mistaken factual beliefs in the public when these beliefs are the result of motivated reasoning in defense of controversial doctrines, and whether factual beliefs, even when supported by our best science, are excluded from public reason if they are entangled in a cultural dispute.

Resumé

Afhandlingen udforsker implikationerne af motiveret ræsonnering, tendensen til at ræsonnering fortolker evidens på måder, der støtter en ønsket konklusion, for normative diskussioner i uenighedens erkendelsesteori og politisk filosofi. Introduktionen giver en kort oversigt over psykologien om motiveret ræsonnering og dens konsekvenser. En af disse er polarisering af overbevisninger: Personer, der er uenige, bliver hver især mere overbeviste, efter de observerer den samme evidens. En anden konsekvens er, at deliberation i grupper med intern uenighed er en effektiv metode til epistemisk forbedring. Introduktionen giver tillige en kort oversigt over uenighedens erkendelsesteori og uenighedens rolle for politisk legitimitet, og påpeger artiklernes bidrag til disse diskussioner.

Artikel 1 svarer på argumenter for den konklusion, at polarisering af overbevisninger er et rationelt fænomen. Der argumenteres for, at forudindtaget behandling af evidens ikke er rationel, når uenighed er fremtrædende, og for at polariserede overbevisninger hverken er dannet på pålidelig vis eller er støttet af evidens.

Artikel 2 diskuterer politisk uenigheds epistemiske betydning. Den viser, at motiveret ræsonnering betyder, at der er en invers korrelation imellem ens opfattelse af en person fra den anden politiske fløjs familiaritet med evidens og deres kognitive evner på den ene side, og deres sandsynlighed for at have ret, hvis man er uenig med dem, på den anden side. Dette generer et problem omkring, hvordan vi skal bestemme sådanne uenigheders epistemiske betydning, som ikke kan undgås ved at benægte, at politisk uenighed har nogen signifikans.

Artikel 3 diskuterer hvilke implikationer de positive effekter af diskussion i grupper med intern uenighed har for uenigheds epistemiske betydning. Der argumenteres for, at disse positive effekter kan generere en epistemisk grund til at holde fast i ens overbevisning, når man opdager uenighed, hvis dette fordrer epistemisk frugtbar diskussion.

Artikel 4 diskuterer, hvilken betydning ræsonnering motiveret af et forsvar for vores politiske og kulturelle værdier har for demokratiske beslutningsprocessers politiske legitimitet. Den adresserer i hvor vidt omfang demokratisk valgte autoriteter bør være lydhøre over for fejltagne faktuelle formodninger, når disse faktuelle formodninger er et resultat af motiveret ræsonnering i forsvar for kontroversielle kulturelle værdier. Den diskuterer også, hvorvidt faktuelle formodninger, der reflekterer ekspertkonsensus og vores bedste videnskab, bør ekskluderes fra den offentlige fornuft, når de er viklet ind i kampe om kontroversielle kulturelle værdier.

Article overview

Article 1: Belief polarization and congeniality bias in reasoning

Individuals tend to construe evidence congenial to their prior belief or desired conclusion as superior to uncongenial evidence. When two people who disagree observe the same evidence, this can result in belief polarization. It has recently been argued that such biased evaluations of evidence, and the resulting belief polarization, are not a sign of irrationality. This article responds to such arguments. I argue that while evaluating evidence in ways congenial to one's prior belief may sometimes be rational, the justifications that have been offered for this conclusion fail in standard cases of belief polarization. Furthermore, the empirical assumption underlying these justifications, namely that congeniality bias in evaluations of evidence is exclusively due to prior belief distributions, is implausible in typical cases of belief polarization. With the exception of some trivial, hypothetical, or artificial cases, belief polarization is not rational.

Article 2: The epistemic significance of political disagreement

The epistemic impact of disagreement is typically thought to be a function of our beliefs about 1) our interlocutor's familiarity with the relevant evidence and arguments, and their intellectual capacities and virtues, relative to our own, or 2) the expected probability of our interlocutor being correct, conditional on our disagreeing. While these two factors are typically used interchangeably, I show that they have an inverse correlation in cases of disagreement about politically divisive propositions. This presents us with a puzzle about the epistemic impact of disagreement in these cases. The most significant disagreements on 1) are the least significant disagreements on 2), and vice versa. I argue that this puzzle cannot be escaped by claiming that we usually have dispute-independent reason to reject the significance of politically charged disagreement.

Published in Philosophical Studies, 2018, May 17, https://doi.org/10.1007/s11098-018-1121-8.

Article 3: Disagreement and the division of epistemic labor

In this article we discuss what we call the deliberative division of epistemic labor. We present evidence that the human tendency to engage in motivated reasoning in defense of our beliefs can facilitate the occurrence of divisions of epistemic labor in deliberations among people who disagree. We further present evidence that these divisions of epistemic labor tend to promote beliefs that are better supported by the evidence. We show that promotion of these epistemic benefits stands in tension with what extant theories in epistemology take rationality to require in cases of disagreement. We argue that the epistemic benefits that result from the deliberative division of epistemic labor can provide epistemic reason to maintain confidence in cases of disagreement. We then show that the deliberative division of epistemic labor constitutes a distinct kind of epistemic dependence.

This article is coauthored with Klemens Kappel and published in Synthese, 2018, April 25, https://doi.org/10.1007/s11229-018-1788-6.

Article 4: Democratic decision-making and the psychology of risk

In many cases, the public (or large parts of it) want to restrict an activity or technology that they believe to be dangerous, but that scientific experts believe to be safe. There is thus a tension between respecting the preferences of the people and making policy based on our best scientific knowledge. Deciding how to make policy in the light of this tension requires an understanding of why citizens sometimes disagree with the experts on what is risky and what is safe. In this paper, we examine two highly influential theories of how people form beliefs about risks: the theory that risk beliefs are errors caused by bounded rationality and the theory that such beliefs are part and parcel of people's core value systems. We then discuss the implications of the psychological theories for questions regarding liberal-democratic decision-making: (1) Should policy be responsive to the preferences of citizens in the domain of risk regulation? (2) What risk-regulation policies are legitimate? (3) How should liberal-democratic deliberation be structured?

This article is coauthored with Andreas Christiansen. I have made some minor modifications to the published version, which can be found in The Ethics Forum, 2017, 12(121), 51-83.

Introduction

1 Introduction

Disagreement is a ubiquitous phenomenon. Individuals disagree about questions of aesthetics, morality, and facts. They disagree about matters that range from the simplistic to the profound, and from the practically insignificant to topics of great importance to human prosperity.

A great deal of philosophical work has sought to understand the normative significance of disagreement. In social epistemology, the most prominent discussion has addressed what (if any) impact evidence of disagreement has on the epistemic rationality of the disputed belief. If an individual believes a proposition and then discovers that someone disagrees, can the very fact of disagreement itself mean that rationality demands a reduction of confidence, or does evidence of disagreement not have this force? For what reasons, and under what circumstances, does disagreement have epistemic significance in this way, if indeed it does? In political philosophy, a central question is what impact widespread and persistent disagreement about values and policy-relevant facts has on the viability and legitimacy of decision-making procedures, institutions, and policies.

Concurrently with these philosophical debates, empirical work in fields such as psychology and political science has sought to understand the psychological mechanisms that cause and maintain disagreements, as well as the consequences of disagreement for group deliberation and problem solving. This research has uncovered a prominent cause of persistent disagreement in the fact that individuals' perception, attention, memory, and reasoning tend to operate in ways that are congenial to prior beliefs or desired conclusions. The presence of disagreement in deliberating groups has been identified as beneficial to the ability of such groups to properly respond to evidence and solve problems.

The four articles in this thesis address aspects of the normative questions about disagreement with a close eye toward the empirical research. They look at what, if any, consequences the picture of human cognition and social interaction that emerges from cognitive, behavioral, and social science has for the arguments and conclusions in the normative debates. In addition, they address some completely novel normative questions that arise from awareness of such research. Articles 1, 2, and 3 discuss epistemological questions about disagreement, while article 4

discusses the implications of factual disagreements and their underlying psychology for the political legitimacy of democratic decision-making procedures and specific policies.

The focus of Article 1 is a normative question that arises out of the tendency for cognition to interpret evidence in ways that are congenial to prior beliefs or desired conclusions. This tendency can cause the beliefs of individuals who disagree to further polarize in response to the same body of subsequently encountered evidence. The article discusses whether such belief polarization, and the biased evaluations of evidence that cause it, can be epistemically rational. It challenges arguments from both traditional epistemology and formal Bayesian models for the conclusion that polarized beliefs in standard cases of belief polarization are rational.

Article 2 turns to the standard question in the epistemology of disagreement, about what degree of belief revision evidence of disagreement requires. Commonly, the answer to this question is thought to depend one of two factors thought to be roughly interchangeable: 1) our interlocutor's degree of familiarity with the evidence and their intelligence, open-mindedness, diligence, etc., relative to our own; 2) our prior subjective probability that they would be right, conditional on our disagreeing. The article shows that motivated reasoning about politically controversial propositions has the puzzling implication that these two factors are inversely correlated in cases of disagreement about such propositions. The more familiar with the evidence and intellectually formidable you think a person on the other side of the political aisle is, the more likely you should think it is that he or she is wrong. This results in a puzzle about how we should determine the epistemic significance of such disagreements.

Article 3 also discusses the question of what, if any, doxastic revision is rationally required when an individual discovers disagreement. It argues that it can be rational to maintain belief in the face of discovered disagreement when the discovery is followed by deliberation with one's interlocutor. The argument proceeds by appeal to empirical research showing that disagreement has beneficial effects on the ability of members of deliberating groups to adopt the belief best supported by the available evidence, and on a defense of a version of epistemic teleology.

Together, articles 1, 2, and 3 suggest that the very same psychological mechanisms that can cause beliefs to be irrational when we individually evaluate evidence about matters that are subject to disagreement can promote rationality when we reason in collectives composed of individuals who disagree.

Article 4 turns to the implications of persistent disagreement about politically charged facts for political philosophy. It presents a tension between the ideal that policy-making should be responsive to the public's preferences and the ideal that policy should be based on our best understanding of the relevant facts, which arises when the preferences of a substantial proportion

of citizens are based in part on factual beliefs that experts dispute. Specifically, citizens' preferences about potential sources of risk may rest on factual beliefs about risk that are a result of their cultural commitments and values biasing their processing of risk-relevant evidence. The article discusses the implications of this psychological picture for the requirement that policy be responsive to citizens' preferences in the domain of risk regulation, for what policies are legitimate, and for how democratic deliberation should be structured if we want citizens to approach an accurate perception of risk-relevant facts.

My main aim in this introduction is to highlight the contributions that the articles make to the normative discussions of the significance of disagreement in social epistemology and political philosophy. Article 1 addresses a normative question that has arisen specifically as a result of psychological research on belief polarization. Since my other contributions to the normative discussions are to a large extent also informed by empirical research, both on this topic as well as the topics of motivated reasoning, confirmation bias, and collective reasoning, I begin by providing an overview of this research, in particular as it relates to disagreement and the arguments of the articles. In light of this empirical background, I then situate the articles within the general literature in the epistemology of disagreement (for articles 1, 2, and 3), as well as the literature on political disagreement and public reason (for article 4), with an eye toward other studies that have integrated psychological findings into their normative discussions. I will also outline the connections between the four articles more clearly, particularly elaborating on the relationship between the epistemic and political dimensions of disagreement. Finally, I discuss some of the theoretical and methodological promises and potential pitfalls of introducing psychological details based on relevant empirical research to debates that are normally conducted in a highly abstract and idealized manner. And of course, no self-respecting dissertation drawing heavily on research about confirmation bias and motivated reasoning can do without some reflections about any possible impact the operation of these biases in myself may have had on the arguments and conclusions herein.

2 Directional biases in reasoning and disagreement

You do not need much familiarity with social science to know that disagreement is pervasive in society. People disagree about a great many things, sometimes vehemently so, and sometimes seemingly in the face of evidence that otherwise appears to settle the matter conclusively.

Psychologists and political scientists have attempted to provide some answers to questions like how disagreements arise on the basis of the same publically available pool of evidence, and how they might be diminished, maintained, or exacerbated. They have studied disagreements about questions in the full range from triviality to profundity: from perceptual disagreements about which of three arbitrary lines is longer (Asch, 1956), to disagreements about the question of whether the theory of evolution by natural selection is true, or about whether childhood vaccines are safe (Kahan & Stanovich, 2016; Nyhan, Reifler, Richey, & Freed, 2014). This research has uncovered a range of social and psychological variables that predict beliefs about hotly disputed questions, such as global warming, gun control, drug policy, health care provision, social security, and much more. These variables include relatively high-level phenomena such as political ideology, group identity, or cultural commitments (Cohen et al., 2007; Dawson, Gilovich, & Regan, 2002; Jost, Nosek, & Gosling, 2008; Kahan, 2015), but also lower-level phenomena such as one's degree of tendency to focus attention on negative stimuli, degree of aversion to novelty, as well as genetics and developmental influences (Funk et al., 2013; Hibbing, Smith, & Alford, 2014; Jost, Glaser, Kruglanski, & Sulloway, 2003).

The line of research that we focus on here, however, has investigated how information processing affects the maintenance of disagreement. When individuals disagree, do they differ in the way that they approach information about the disputed topic? If so, what consequences does that have for the dispute? What consequences does disagreement have for our ability to arrive at beliefs that are supported by our publically available evidence, in both small collectives and society at large? This research is broadly relevant to all four articles, but in what follows I will highlight when and how an article makes particular reference to the research.

2.1 Directional goals and accuracy goals

It is sometimes almost considered a truism that the function of reasoning, and of cognition more generally, is to arrive at accurate representations of the world. A common assumption is that the reason evolution has furnished us with sophisticated and metabolically costly information

processing capacities is that they improve our ability to act appropriately by yielding accurate representations of the environment (Inhelder & Piaget, 1958; Stanovich & West, 2003). This assumption is widespread even in those areas of psychology that have produced evidence of ubiquitous irrationality in human cognition, such as the heuristics and biases literature or the psychology of deductive reasoning (Evans, 2002; Kahneman, 2003). Such irrationality is taken as performance errors, due to limitations in our cognitive capacities, rather than as expressions of irrationality being, at some level, functional.

While this picture is most likely accurate most of the time, cognition does not *always* function so as to maximize the correspondence between our mental representations and the world within the limits set by our cognitive capacities (Kahan, 2017; Mercier & Sperber, 2011). There are occasions where we process information not with the aim of arriving at the most accurate interpretation of the evidence, but at an interpretation that supports a conclusion that we find desirable (Kunda, 1990).

Kunda (1990) distinguishes between accuracy goals and directional goals in cognition. When cognition is driven by accuracy goals, it aims at arriving at an accurate assessment of the evidence, and ultimately at accurate beliefs about the world. When cognition is driven by directional goals, it (unbeknownst to the subject) aims at arriving at a construal of evidence that allows for the subject to reach a desired conclusion while maintaining an "illusion of objectivity" – an image of one self as an evidence-driven, objective believer.

The goals in question can vary. A widely discussed directional goal is a desire to confirm or defend a currently held belief. When such a goal is in play, we say that the resulting information processing is subject to *confirmation bias* (Nickerson, 1998). But directional goals in information processing can derive from other things than prior belief. Examples include our situation-specific practical goals (Balcetis & Dunning, 2006; DeScioli, Massenkoff, Shaw, Petersen, & Kurzban, 2014), defense of our personal or social identity (Nel & Steele, 2000; Sherman & Cohen, 2002; Sherman, Kinias, Major, Kim, & Prenovost, 2007), and a desire to arrive at beliefs that are commonly accepted within one's affinity groups (Kahan, 2017; Kahan, Jenkins-Smith, & Braman, 2011). In these cases we call the resulting information processing *motivated cognition,* and, in the special case of reasoning, *motivated reasoning*.

The question of the balance between accuracy goals and directional goals in reasoning is touched upon at some length in articles 1 and 4. Article 1 shows that prominent arguments for the rationality of belief polarization (which will be described below) rely on the assumption that accuracy goals are behind confirmation bias in reasoning. It is true that in some circumstances,

confirmation bias can derive from accuracy goals. If one reasonably suspects that evidence against one's prior belief must somehow be flawed, then an accuracy goal can spur one to selectively scrutinize this evidence in order to locate the flaws, and thereby arrive at what one suspects is the accurate assessment of the evidence. Selectively scrutinizing evidence according to whether it agrees with your prior belief is a form of confirmation bias, but one that, in this case, is driven not by a desire to defend one's prior, but a suspicion that a correct assessment of counterevidence requires that it receives extra scrutiny. Article 1 argues that such models based on accuracy motivation are inadequate in cases of belief polarization. Rather than being a result of accuracy-driven confirmation bias, belief polarization as demonstrated in the literature is the result of motivated reasoning, and this has negative implications for our assessment of its rational status.

Article 4 discusses the balance of accuracy goals and directional goals in cognition about risk. It contrasts a so-called bounded rationality model based on the notion that cognition has accuracy goals, but errors can occur due to a lack of evidence or cognitive capacity (Kahneman, 2003; Sunstein, 2005), with one that also includes motivated reasoning in defense of one's cultural commitments (Kahan, Braman, Gastil, Slovic, & Mertz, 2007), as explanations of disagreements between experts and substantial portions of the public. It argues that the motivated reasoning model provides a better explanation of the politically clustered and divided nature of beliefs about politically salient sources of risk, and discusses the normative political implications of this view.

2.2 Motivated reasoning

Directional goals can be implemented at several levels of cognition. At the lowest level, they can influence sensory perception: Balcetis and Dunning (2006), for example, found that subjects' current goals influenced how they perceived ambiguous visual stimuli (e.g. whether a figure was perceived to be a B vs. a 13; or a horse vs. a seal). Directional goals also affect how we seek out and attend to information. In particular, subjects overwhelmingly tend to seek out and attend to information that is congenial to their desired conclusion, even when experimenters instruct them to try to be objective, or when they are given monetary incentives to expose themselves to evidence for the opposing view (Frimer, Skitka, & Motyl, 2017; Hart et al., 2009; Jones & Sugden, 2001; Taber & Lodge, 2006). With respect to memory, we are more likely to recall evidence that supports a desired conclusion and to forget evidence that tells against it. We are furthermore likely to misremember information as being more supportive of our desired

conclusion than it is (Hennes, Ruisch, Feygina, Monteiro, & Jost, 2016). Finally, directional goals can bias our reasoning. When reasoning about some matter, we tend to spontaneously produce reasons in favor of our desired conclusions, and not reasons against it (Koriat, Lichtenstein, & Fischhoff, 1980; Mercier & Sperber, 2011; Petersen, Skov, Serritzlew, & Ramsøy, 2012). When we are asked to evaluate reasons, whether they take the form of arguments, statistical data, or something else, we tend to be much more critical of reasons against our desired conclusion than we are of reasons in its favor. Indeed, we often spontaneously produce potential defeaters of these reasons, and additional counterarguments in favor of our desired conclusion, when faced with reasons against a desired conclusion. We spend much longer evaluating reasons if they tell against a desired conclusion, and spend this time denigrating the reasons. The more limited time that we spend evaluating reasons in favor of our view is used to praise them as being eminently good reasons to believe the desired conclusion. The result is that reasons for our desired conclusion tend to be considered much stronger than reasons against it (Dawson et al., 2002; Kahan, 2016; Kraft, Lodge, & Taber, 2015; Taber, Cann, & Kucsova, 2009; Taber & Lodge, 2006).

Motivated reasoning has received surprisingly little sustained philosophical attention. Although it is entirely possible that I have missed some relevant entries to the literature, my search found few detailed discussions of the epistemic significance of these findings. I am aware only of articles by Kornblith (1999), Kenyon (2014), Ballantyne (2015), Jern et al. (2014), Boudry & Braeckman (2012), and Kelly (2008). In political philosophy, works by Richey (2012), Kahan (2007; 2006), Landemore (2012; Mercier & Landemore, 2012), Sunstein (2006), and Bagg (2015), have addressed the impact that motivated reasoning has on the viability of deliberative democracy and the legitimacy of institutions and policies.

The four articles in this thesis thus enter into what I think are philosophical discussions that deserve much more attention than they have been given thus far.

2.3 Belief polarization and cultural cognition

What do these psychological mechanisms mean for disagreement? Consider scenarios where two people disagree about some proposition, and they each have directional goals to defend their belief. They subsequently encounter the same body of evidence pertinent to the disputed proposition. Each of them will be inclined to process the evidence in a manner that is congenial to their prior belief in the ways described above. Perhaps they are each successful enough in this

that they arrive at a construal of the evidence that they take to support their desired conclusion. So they both strengthen their view, resulting in the magnitude of the disagreement increasing.

This phenomenon is called *belief polarization* – subjects' beliefs move closer toward the poles of absolute uncertainty and absolute certainty with respect to the disputed proposition.[1] The fact that they do so after subjects observe *the same evidence* suggests that polarization occurs due to the impact of confirmation bias or motivated cognition.

Belief polarization, in the narrow sense that beliefs polarize after subjects who disagree observe the same evidence, has been observed for topics including the death penalty, religious belief, beliefs about homosexuals, global warming, gun control, political candidates, abortion, environmental protection, nanotechnology, the HPV vaccine, and several others (Batson, 1975; Cook & Lewandowsky, 2016; Kahan, Braman, Cohen, Gastil, & Slovic, 2010; Kahan, Braman, Slovic, Gastil, & Cohen, 2009; Lord, Ross, & Lepper, 1979; Miller, McHoskey, Bane, & Dowd, 1993; Munro & Ditto, 1997; Munro, Ditto, Lockhart, Fagerlin, & Gready, 2002; Pomerantz et al., 1995; Wilson, Kraft, & Dunn, 1989). What these topics have in common is that they tie into personal, social, political, and cultural identities in ways that are likely to trigger directionally motivated reasoning.

Article 1 directly discusses the epistemic rationality of belief polarization, in the narrow sense of polarization following exposure to identical evidence. It argues, against views put forth by Kelly (2008) and Jern et al. (2014) to the contrary, that confirmation bias and motivated reasoning are not rational when one evaluates evidence about a proposition that is known to be the subject of disagreement. As a result, belief polarization that results from such biases is not epistemically rational.

There is another, broader, notion of belief polarization. One often hears that beliefs about some issue in society are polarized or have polarized over time. What this typically means is that beliefs are polarized along political, ideological, or cultural fault lines, such that, for example, a person's position on the political spectrum is highly predictive of the direction and extremity of their doxastic attitude toward the proposition. For example, partisanship and political ideology is highly predictive of beliefs about anthropogenic global warming. Beliefs about this issue are polarized along political lines, and the degree of polarization has increased since the 1990s

[1] It is more accurate to say that philosophers call the phenomenon belief polarization. In the psychological literature, it is typically referred to as attitude polarization.

(Kahan et al., 2012; McCright & Dunlap, 2011; McCright, Xiao, & Dunlap, 2014; Pew Research Center, 2016; Shi, Visschers, & Siegrist, 2015).

An interesting finding with respect to such polarization is that *correlates positively* with measures of cognitive ability, education, scientific literacy, and even intellectual virtues such as open-mindedness (Hamilton, 2011; Kahan et al., 2012; Kahan & Corbin, 2016; Kahan & Stanovich, 2016). The more intelligent, reflective, open-minded, or well-educated a person is, the more likely they are to adopt an extreme belief about the disputed issue, and this goes for both sides of the dispute. So, for example, liberal democrats, or people with a communitarian and egalitarian cultural outlook, tend to believe that humans are causing global warming, and their certainty of this increases with their level of education, cognitive ability, open-mindedness, scientific literacy, etc. In contrast, conservative republicans, or people with a hierarchical and individualist cultural outlook, tend to disbelieve that humans are causing global warming, and their certainty that we are not increases with their level of education, cognitive ability, open-mindedness, scientific literacy, etc. (Kahan et al., 2012).

This might strike some people as very surprising. After all, these abilities ought to increase the likelihood that one is able to arrive at the doxastic attitude that is best supported by the publically available evidence. It seems that people with greater ability and intellectual virtue ought to be the most likely to converge on the view that is supported by our best science. However, according to the *cultural cognition thesis*, the observed pattern is perfectly explicable by motivated reasoning in defense of people's cultural values (Kahan, 2012). The question of anthropogenic climate change has, like several other matters of fact, become embedded in a broader cultural and political struggle about how we should arrange society: To what extent is human flourishing best promoted by individually-driven, free-market, bottom-up solutions to societal problems, as opposed to collectively-driven, top-down solutions? To what extent is it best promoted by hierarchical social structures with clear status- and power differentials, as opposed to flat, egalitarian ones? In the current political climate, at least of countries including the U.S., Australia, Switzerland, the UK, and Norway, the reality of anthropogenic global warming is perceived as vindicating the collectivist and egalitarian side of this dispute (Cook & Lewandowsky, 2016; Kahan, Silva, Tarantola, Jenkins-Smith, & Braman, 2015; Shi et al., 2015; Aasen, 2017). Its reality would impugn the ability of individuals, corporations, and societal elites to properly manage challenges facing society, and would suggest the need for collectively imposed top-down controls on their activity. Due to this perception, the prospect of believing that anthropogenic global warming is occurring is threatening to the cultural values of

hierarchical individualists and free-market supporters, whereas egalitarian communitarians see it as a vindication (Heath & Gifford, 2006; Kahan et al., 2015; Lewandowsky, Gignac, & Oberauer, 2013). This entanglement of facts about global warming and cultural values generates a directional goal for cognition to construe the evidence as supportive of the factual conclusion that is congenial to one's cultural values. Those with the most information and greatest cognitive capacities are the most likely to succeed in this goal: They have more resources with which to generate arguments in favor of their view, to find supposed flaws in the arguments against their view, and in general to rationalize the congenial conclusion. Indeed, experiments show that those with the greatest cognitive capacities are, in a certain sense, *more* biased in their processing of evidence than the less well endowed. Not because their directional goal is stronger, but because they are better equipped at recognizing how to make the evidence yield the desired conclusion while maintaining the "illusion of objectivity" (Kahan, 2013; Kahan, Peters, Dawson, & Slovic, 2017).

Articles 2 and 4 discuss the implications of this pattern. Article 2 suggests that the observed correlation between cognitive ability and polarization generates a puzzle in the epistemology of disagreement. Here, two notions of what generates epistemic reason to revise one's doxastic attitude in cases of disagreement are employed: (1) the other person's familiarity with relevant evidence and their general abilities at processing it, relative to one's own; (2) one's subjective probability (prior to discovering the disagreement) that the other person is right, conditional on their disagreeing. (1) and (2) are usually taken to be more or less coextensive, but article 2 shows that, in cases with the observed pattern between ability and polarization, (1) and (2) are inversely correlated. This forces a choice between the two in how to determine the epistemic significance of disagreement, but both options have puzzling implications. Article 4 discusses the consequences of the distribution of factual beliefs for the political legitimacy of policies about domains such as global warming, where cultural cognition is involved in shaping the factual beliefs of the citizenry. Supposing that the best scientific evidence really does support that anthropogenic global warming is occurring and is a risk to human prosperity, and that experts have reached consensus or near-consensus about this, how should democratic states respond to the presence of disagreement in a large proportion of the public? To what extent is it necessary for the democratic and liberal legitimacy of decision-making procedures and policies that they are responsive to the factual dissent, when it is based on what we know to be a biased evaluation of the evidence that results from their controversial cultural commitments?

2.4 Collective reasoning and disagreement

The cited psychological findings may paint a rather bleak picture of human reasoning and disagreement. However, other research has shown that disagreement can be a source of epistemic boons in addition to maladies. In particular, this result comes from research on reasoning in collectives. When groups reason collectively (that is, when members exchange reasons and arguments), the presence of disagreement within the group increases the likelihood that evidence for both sides of the issue is given a proper hearing, that groups do not prematurely settle on a suboptimal solution, and that the group ultimately arrives at the conclusion best supported by the available evidence.

Recall that individuals tend to spontaneously generate arguments for their views and counterarguments against challenges. When they reason in isolation, they further tend to be cognitive misers: they expend only what effort on generating these reasons is necessary for them to maintain the illusion of objectivity (Toplak, West, & Stanovich, 2014; Trouche, Johansson, Hall, & Mercier, 2016). When they reason in collectives, however, they anticipate that their reasons will be challenged. This increases their motivation to expend cognitive resources on coming up with good reasons, and indeed, the reasons that people generate in dialogic settings, or in anticipation of such settings, tend to be much better than the reasons they generate when alone (Kuhn, Shaw, & Felton, 1997; Mercier & Sperber, 2011). In addition, in spite of the tendency for motivated reasoning to judge argument strength according to congeniality with the desired belief, subjects in collective deliberations do in fact change their minds in response to strong argumentation (Petty, Wegener, & Fabrigar, 1997; Vinokur & Burnstein, 1978).

The effect of this is that deliberating groups with internal disagreement will often have a deep pool of reasons available for both sides of the issue, and the capacity to sort of the good reasons from the bad. As a result, they tend to vastly outperform both individuals and less diverse groups on a variety of cognitive tasks. This includes solving deductive and mathematical problems, finding the optimal solution in a solution space with several local optima but only a single global optimum, finding creative solutions to problems requiring some sort of insight, finding the conclusion best supported by the group's total evidence when the evidence is only partially shared, and measures of the performance of work groups in natural settings (Baumeister, Ainsworth, & Vohs, 2016; Duarte et al., 2015; Hong & Page, 2004; Laughlin & Ellis, 1986; Mayo-Wilson, Zollman, & Danks, 2013; Mercier, Deguchi, Van der Henst, & Yama, 2015; Mercier, Trouche, Yama, Heintz, & Girotto, 2015; Michaelsen, Watson, & Black, 1989; Moshman & Geil, 1998; Muldoon, 2013; Polzer, Milton, & Swann, 2002; Schulz-Hardt, Brodbeck, Mojzisch, Kerschreiter, & Frey, 2006; Schulz-Hardt, Jochims, & Frey, 2002; Trouche

et al., 2016; Trouche, Sander, & Mercier, 2014; van Knippenberg & Schippers, 2007; Watson, Kamalesh, & Michaelsen, 2016; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010; Woolley, Aggarwal, & Malone, 2015).

Contrast this with groups that lack internal disagreement. In this case, individuals' motivated assessment of the evidence is likely to be reinforced by the presence of others with the same desired conclusion. Members spontaneously tend to generate reasons pointing in the same direction, and to be congenial in their assessments of these reasons. Some of the reasons will be novel to some participants, giving them even more reason to believe their desired conclusion. Members are unlikely to share any information they might have that counts against the favored conclusion, and are liable to immediately conclude that a desired conclusion is correct rather than search for alternatives. Famously, such groups are vulnerable to *group polarization*. Group polarization is the term used for the phenomenon that the average credence of a deliberating group with an initial inclination moves toward the extreme in the direction predicted by the initial inclination after deliberation (Isenberg, 1986; Moscovici & Zavalloni, 1969; Sunstein, 2002b). While there might be cases where group polarization is rational (after all, it has been argued, it can lead to certainty of a truth (Easwaran, Fenton-glynn, Hitchcock, & Velasco, 2016)), common opinion has it that it is an unfortunate consequence of group deliberation. In groups with a sufficient diversity of views, or that decompose into two subgroups that disagree, one tends instead to find *depolarization*: the distance between the credences of the members, or between the two subgroups, tends to diminish due to the voicing of diverse arguments that are novel and of a generally high quality (Abrams, Wetherell, Cochrane, Hogg, & Turner, 1990; Vinokur & Burnstein, 1978).

Article 3 relies heavily on the benefits of disagreement to deliberating groups for its conclusion. It argues that their existence provides one with an epistemic reason not to reduce confidence based on the evidence that one is party to a disagreement. Reducing confidence would preclude the beneficial epistemic effects of disagreement, and increase the risk of the pernicious effects associated with prior agreement, to any ensuing group deliberation. Combining the empirical results with a notion of epistemic rationality that allows for epistemic teleological concerns to figure in normative evaluation, epistemic rationality would not require a reduction of confidence, even if a reduction of confidence is what one's evidence supports at the time the disagreement is discovered.

3 The epistemic significance of disagreement

3.1 Disagreement in epistemology

Two people disagree about a proposition p when they adopt different doxastic attitudes toward p. On a tripartite view, doxastic attitudes are belief, disbelief, or suspension of judgment. Most standardly, two people disagree if one believes p and the other disbelieves p. But they also disagree if one (dis)believes p and the other suspends judgment. On a subjective probability view, two people disagree if they adopt different credences toward p. If I believe p to degree .2 and you believe p to degree .8, we disagree, but we also disagree, in some sense, if I believe p to degree .97 and you believe p to degree .98.

The *core question* in the epistemology of disagreement is, as I take it, what (if any) effect evidence that one is party to a disagreement has on the epistemic rationality or justification of one's doxastic attitudes. The debate surrounding the core question has, to a great extent, revolved around judgments about cases. So let us begin by considering a few such cases of disagreement:

NEIL: I believe that Saturn is the only planet in our solar system with rings. I then strike up a conversation with the astrophysicist Neil. As the conversation goes on it becomes apparent that he believes that there are four planets with rings in our solar system.

RESTAURANT: Five of us go out to dinner. It's time to pay the check, so we're interested in how much we each owe. We can all see the bill total clearly, we all agree to give a 20 percent tip, and we further agree to split the whole cost evenly. I do the math in my head and become highly confident that our shares are $43 each. Meanwhile, my friend does the math in her head and becomes highly confident that our shares are $45.

BUS STOP: While waiting at the bus stop, I am approached by Bojan, who tells me that he is certain I am living in a shoe. I am fairly confident, based on long familiarity with my apartment, that I live in an apartment, not a shoe.[2]

[2] NEIL comes from Matheson (2015, p. 19), RESTAURANT from Christensen (2007, p. 193), and BUS STOP from Weatherson (2016, p. 214), all with minor changes in formulation.

The three cases evoke different judgments about whether one should reduce confidence in response to discovering the disagreement. In the case of NEIL, the correct verdict seems to be that one should defer to Neil's judgment. In contrast, there seems to be no real pressure to reduce confidence that one lives in an apartment in BUS STOP. RESTAURANT presents an intermediate case, where some reduction of confidence seems required (although this conclusion is not universally agreed upon). Cases like NEIL and BUS STOP are not very informative about the epistemic significance of disagreement as such. Neil is an expert about the matter at hand, and I am not, so any reason for me to defer could simply result from my recognition of his superior epistemic position, rather than from the significance of disagreement itself. Likewise, I have good reason to suspect that I have more probative evidence about my own living arrangements than Bojan does, so my lack of a reason to reduce confidence could be the result of my awareness of being in a much better epistemic position than he is.

In order to arrive at a more precise investigation of the epistemic significance of disagreement, philosophers have largely limited their discussions to cases that are idealized in a number of ways. The most important idealization is that the interlocutors are assumed to be epistemic peers. According to Kelly (2005, p. 174), who first introduced the notion of epistemic peerhood to the modern debate, two individuals are epistemic peers with respect to a question if and only if they satisfy the following two conditions:

(i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and
(ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias.

What matters for the purposes of epistemic evaluation is typically taken to be whether I am justified in believing that you are an epistemic peer in this sense, not necessarily that this is in fact the case. As I understand the notion of equality at work in (i) an (ii), being equals need not mean being identical. To be evidential equals implies that the individuals have an equally good familiarity with relevant evidence and arguments, not necessarily that the evidence and arguments that one is familiar with are exactly the same as the other person is familiar with (Christensen, 2007; cf. King, 2012). That two individuals are equals with respect to epistemic virtues can be roughly paraphrased as saying that the two individuals are equally good at processing evidence, not necessarily that they do so in the same way. Suppose, for example, that

two individuals disagree and are both subject to motivated reasoning about the disputed proposition. In their processing of the evidence pertinent to the disagreement, they are disposed to reach contrary evaluations of its bearing on the disputed proposition. Nevertheless, they may be equals in the sense of (ii) if they are biased to an equal extent.

Elga (2007, p. 499) offers a slightly different notion of epistemic peerhood. On his view, you count someone as your epistemic peer "…with respect to an about-to-be-judged claim if and only if you think that, conditional the two of you disagreeing about the claim, the two of you are equally likely to be mistaken."

While different, the two notions are related. How likely you think someone is to be mistaken about a claim relative to yourself depends on your beliefs about how his or her familiarity with the evidence, and his or her virtues relevant to the processing of the evidence, stack up against your own. However it is construed, the condition of epistemic peerhood is meant to ensure that asymmetries in epistemic position are not behind judgments about rationality in the target cases. The greater symmetry in cases of peer disagreement increases the likelihood that a requirement to alter one's doxastic state is due to the disagreement as such.

In article 2, I argue that, appearances to the contrary notwithstanding, one's choice of criteria for peerhood can have dramatic consequences for the evaluation of the epistemic significance of disagreement. In disagreements where motivated reasoning plays a prominent part, such as politically structured disagreements, our evidence and our intellectual virtues are put to use to defend the politically congenial conclusion. I show that this implies that in a political disagreement, I should think it more likely that you are wrong if you have great familiarity with the evidence and arguments and are intellectually virtuous in the sense of (ii). The two notions do not only come apart, they are inversely correlated.

Another idealization employed for the purpose of eliminating any potential asymmetries is full disclosure. Full disclosure obtains when the parties to a disagreement "…have thoroughly discussed the issues. They know each other's reasons and arguments, and that the other person has come to a competing conclusion after examining the same information" (Feldman, 2006, p. 419). Like epistemic peerhood, full disclosure is meant to erase factors other than the disagreement itself from the equation when evaluating the significance of disagreement. In the absence of full disclosure, the possibility that the other person has evidence or arguments available that one does not might be what provides any reason to reduce confidence.

Articles 2 and 3 engage with the core question, but they deal with real-life cases of disagreement that presumably fall well short of epistemic peerhood and full disclosure obtaining. However, this does not mean that these articles are talking at cross-purposes with extant discussion of disagreement in epistemology based on idealized cases. The ubiquity of idealizations in the epistemology of disagreement does not mean that epistemologists take more ordinary disagreements to be epistemically insignificant (Christensen, 2014b; Lackey, 2008b; Matheson, 2015). While this is not always articulated in the literature, idealizations should, at least to my mind, be viewed primarily a methodological tool rather than a substantive view about the conditions under which evidence of disagreement imparts rational requirements on our doxastic attitudes. As I see it, idealizations serve the dual roles of providing the discussion with clarity and playing a similar role to that of experimental control of confounding variables in science. Scientists want to control for confounding variables because only by successfully doing so are they warranted in inferring that the independent variable was what caused an observed change in the dependent variable. But this does not imply that they think that the independent variable plays no role outside of the context of the experiment. Similarly, idealizations in the epistemology of disagreement allow for control of what are, for the purposes of investigating the epistemic significance of disagreement as such, considered to be confounding factors. This allows inferences about the impact of disagreement on the rationality of doxastic attitudes, but does not imply that disagreement is epistemically insignificant outside of this controlled context.

3.2 The main responses

Weatherson (2016) quips that there are, as usual in philosophy, slightly more answers to the core question than there are philosophers working on it. Fortunately, there is a useful spectrum along which these many answers are typically placed. Keeping our focus on idealized disagreements for now allows for a clearer statement of the various views. So in the following presentation we will assume that the parties to disagreement are epistemic peers, and that the condition of full disclosure obtains.

Conciliatory views hold that it is rationally required for both parties to an idealized disagreement to reduce confidence, usually significantly so. *Steadfast* views hold that there are cases of idealized disagreement where at least one party to the disagreement is rationally permitted, or even required, to maintain confidence. The next sections present a very selective sample of views from both sides of the spectrum, as well as those hard-to-classify cases around the middle, and some arguments in their favor.

3.2.1 Conciliatory views

Consider RESTAURANT again, with the stipulation that my friend and I are epistemic peers. A conciliatory view would hold that upon discovering our disagreement, it is no longer rational for me, or for my friend, to hold our original doxastic attitude. Instead, we should both rationally adopt a doxastic attitude that is closer to our interlocutor (Christensen, 2007). The fact that my friend, who is my epistemic peer, has reached a different verdict, defeats whatever prima facie justification I had for my original doxastic attitude. This conclusion can, and has been, motivated in various ways. One is to start from the case of expert disagreement. In NEIL, it is clear that I should defer. Neil is an expert, and is very likely to be right about the number of planets in our solar system that has rings. I have a strong reason to think that Neil's answer is correct. But it seems that I also get a reason, albeit less strong, to think that my friend is right in RESTAURANT. After all, my friend is generally good at calculating shares of bills, and good as I am. We can stipulate that we both have track records of being right 99% of the time when calculating shares in this way. It seems that my friend reaching a particular result gives me some reason to think that this result is correct. Another way to motivate the conclusion is to consider a variation where, rather than discovering that one epistemic peer disagrees with me, I find out that 1,000 epistemic peers have, independently of one another, reached a verdict different from mine. It seems clear that I should defer to the majority in this case, or at least become much less confident in my belief. But if the disagreement of 1,000 epistemic peers has a strong impact, surely the disagreement of one epistemic peer has some impact as well (Kelly, 2010; Lackey, 2010).

While conciliatory views agree that evidence of peer disagreement provides a defeater of one's original doxastic attitude, this leaves open the question of how strong this defeater is, and how much doxastic revision is required. It is useful in presenting this view, and for the ensuing discussion in general, to draw a distinction between first-order evidence and higher-order evidence (Christensen, 2010; Kelly, 2010; Matheson, 2009). In RESTAURANT, let's refer to the proposition that *the share is $43* as p, and refer to the bill total, plus the fact that we have agreed to add a 20% tip and split the bill evenly, as E. E is first-order evidence about p – it is evidence that directly bears on the truth of p. *E supports p* is a higher-order proposition with respect to p. Higher-order propositions are about evidential relations, or about one's capacity to grasp evidential relations. The fact that I have inferred that p follows from E, and that I am usually

right about such matters, is higher-order evidence. It is evidence that bears on the truth of a higher-order proposition, in this case the proposition *E supports p*.

On the Equal Weight View (Elga, 2007; Matheson, 2009) I should give as much weight to higher order evidence about my epistemic peer as I give to higher order evidence about myself. Since we are epistemic peers, the fact that my peer has inferred that p does not follow from E is just as strong evidence against *E supports p* as my own having inferred p from E is evidence in its favor. One of us has made some error, and since we are epistemic peers, there is no reason to suspect that it is more likely that it is one rather than the other. The effect is that any propositional justification I may have had for *E supports p* is defeated, and this in turn defeats my justification for believing p (Kappel, 2017). On one interpretation of the Equal Weight View, it mandates *splitting the difference* with respect to p. Given that I give equal weight to the possibility that you are right and the possibility that I am right, and that you should do the same, our doxastic attitudes toward p should meet at the middle of the original gap between our attitudes. If I initially believe that p and you disbelieve it, we should both suspend judgment about p. If my initial credence in p was .8 and yours was .2, then we should converge on .5. If my initial credence was .9 and yours was .4, we should converge on .65.[3]

Elga (2007) offers a defense of the Equal Weight View based on the seemingly absurd consequences of denying it. Consider the following case:

RACE: You and a friend whom you consider to be your epistemic peer are to judge the same contest, a race between Horse A and Horse B. Initially, you think that your friend is as good as you at judging such races. In other words, you think that in case of disagreement about the race, the two of you are equally likely to be mistaken. The race is run, and the two of you form independent judgments. As it happens, you become confident that Horse A won, and your friend becomes equally confident that Horse B won (Elga, 2007, p. 486).

The Equal Weight View would say that you should consider the probability that you are right to be 50% after discovering the disagreement in RACE. Elga (2007, pp. 486–487) argues that any other probability leads to absurdity when considering a long series of such cases: "Suppose that you and your friend independently judge the same long series of races. You are then allowed to compare your friend's judgments to your own. (You are given no outside information about the race outcomes.) Suppose for reductio that in each case of disagreement, you should be 70%

[3] Naturally, the same convergence should occur with respect to the proposition that our share is $45.

confident that you are correct. It follows that over the course of many disagreements, you should end up extremely confident that you have a better track record than your friend. As a result, you should end up extremely confident that you are a better judge. But that is absurd."

The worry is that, since you started out believing that your friend was your epistemic peer, it is doubtful that you could rationally end up believing that you are your friend's clear superior, when you have received absolutely no direct evidence to that effect. But a non-equal weight view seems to entail this.

Suppose that, in RESTAURANT, I was to reason along the following lines: "Before we calculated our shares, I considered you to be my epistemic peer. But since p really does follow from E, and you believe that not-p on the basis of E, I must have been wrong in considering you my peer in this matter. Since it turns out that you're not my peer after all, I should be unmoved by your disagreement about p."

Christensen (2009) argues that such reasoning would be illegitimate. He puts forth the following principle to explain why:

INDEPENDENCE: In evaluating the epistemic credentials of another's expressed belief about p, in order to determine how (or whether) to modify my own belief about p, I should do so in a way that doesn't rely on the reasoning behind my initial belief that p (Christensen, 2009).

INDEPENDENCE says that, in cases like RESTAURANT, I cannot rely on my having concluded p from E in my assessment of your epistemic position. I should, in a certain sense, bracket E and my reasoning about E when determining your epistemic standing with respect to my own. A legitimate reason to downgrade my assessment of your epistemic position needs to be independent of the dispute in question. As stated, there is no such independent reason in RESTAURANT, so I should consider you my peer and decrease my confidence in p accordingly. If, however, I were to note that you had consumed two bottles of wine during our lunch, while I had water, this would give me a dispute-independent reason to think that I am in a better epistemic position, since your intoxication plausibly has a negative impact on your ability to work out our shares.

Attributions of reasoning biases are sometimes given this role of providing dispute-independent reason to downgrade the epistemic position of an interlocutor. If I discover that we disagree, and I think that I can explain your being mistaken by reference to your being under the influence of motivated reasoning, then this could constitute a dispute-independent reason to

downgrade your epistemic credentials. For example, Fumerton (2010, p. 102), in discussing how he ought to respond to disagreements with other philosophers, says as follows: "Perhaps I have some reason to believe, for example, that [my disagreeing colleagues] are the victims of various biases that cause them to believe what they want to believe. Indeed, I suspect that I do have reason to believe that others are afflicted in such ways…" However, Fumerton does not think that the same applies to him: "I do, in fact, think that I have got more self-knowledge than a great many other academics I know, and I think that self-knowledge gives me a better and more neutral perspective on a host of philosophical and political issues" (2010, p. 102). Fumerton thinks that, by attributing motivated reasoning to other philosophers, even well known and respected ones, while relying on introspection to conclude that he is not so afflicted, he is sometimes justified in discounting their epistemic position to some extent when they disagree with him.

In article 2, I discuss this argumentative strategy as it applies to cases of political disagreement. If you find yourself in a political disagreement, it is plausible that your interlocutor's belief is partly the result of motivated reasoning. But this does not give you a reason to denigrate his or her epistemic position relative to your own, because your own belief is equally likely to be so afflicted. Research shows that motivated reasoning is distributed equally among liberals and conservatives (Frimer et al., 2017; Hallen, Bingham, Hill, Carolina, & Cohen, 2017; Kahan, 2013), so the political position of an interlocutor cannot itself generate an asymmetry. Fumerton argues that self-knowledge and introspection allow him to conclude that motivated reasoning does not afflict his own beliefs, but I show that this argument fails, for reasons similar to those presented by Ballantyne (2015). Research on so-called bias blind-spot (Pronin, Lin, & Ross, 2002), a tendency to attribute biases to others but not to oneself, regardless of whether one is in fact biased, and research suggesting that introspection is a poor method of coming to know our own cognitive processes (Carruthers, 2011; Nisbett & Wilson, 1977), implies that we are not justified in attributing motivated reasoning to political opponents but not ourselves on the basis of introspection. So while it might be true that my interlocutor is subject to motivated reasoning, I have no independent reason to suspect that I am not subject to the same, even if introspection seems to suggest that I am not.

Kelly (2010) has objected that INDEPENDENCE means that conciliatory views put implausibly much weight on higher order evidence and too little on the first-order evidence in determining what doxastic attitude is rational. It is implausible that we should 'bracket' our first-order evidence and our reasoning about the first-order evidence when settling a dispute, as

INDEPENDENCE would have it. We return to this when we discuss Kelly's Total Evidence View.

Conciliatory views in general have faced the objection that they lead to widespread skepticism, or, as Elga (2007) calls it, to 'spinelessness'. The worry is that conciliatory views imply that we ought to give up our views about many controversial topics in areas such as politics, philosophy, or science, since we disagree about these issues with people who are presumably our epistemic peers (and with some who are our superiors).

Elga's response to this problem is that in these real-life disagreements, we often take ourselves to have dispute-independent reasons to downgrade the epistemic position of those with whom we disagree. One such reason is that, in real life disagreements in domains like politics, views often form clusters. If I know your position on abortion, I probably also know your position on gun control, marihuana legalization, and global warming (Kahan et al., 2007). When we disagree about abortion, I can draw on your being mistaken (by my lights) about these other, related issues to deny your being my epistemic peer with respect to abortion.

As I argue, again in article 2, this response is not very persuasive. It may very well be descriptively accurate that we tend to downgrade the epistemic position of those with whom we find ourselves in pervasive disagreements. But I do no see why we are epistemically rational in so doing from the perspective of the Equal Weight View. Take the case of politics. It is true that many controversial issues form clusters, and that people's beliefs about one issue are highly predictive of their beliefs about the others in the cluster. As I described in the section on cultural cognition, one empirically well-supported explanation for this is that specific political issues, including factual beliefs relevant to these issues, become embedded in a broader cultural contest. Our own cultural commitments cause us to construe arguments and evidence as supportive of those beliefs that affirm that our cultural outlook is superior (Kahan, 2012, 2015). But an argument that we can downgrade those with whom we disagree with over controversial issues on this basis appears to fall victim to the same kind of objection that Fumerton's (2010) argument did: We may have no dispute-independent reason to think that our beliefs about anything in the cluster are based on any less of a biased processing of the evidence, as people on both sides of the political divide are equally vulnerable to these biases. We also may have no independent reason to think that we are so lucky that our worldview and cultural commitments have inclined us to have true beliefs that the evidence really does support, while the other side's cultural outlook has inclined it to false beliefs. That is not to say that we never have independent reasons to downgrade the epistemic standing of an interlocutor for particular politically divisive beliefs.

Article 2 cites the fact that the vast majority of experts agree that humans are causing global warming, and that such a verdict correlates with level of expertise in climate science (Cook et al., 2013, 2016), as an independent reason to think that one side of this disagreement is right. One might also point to the evidence of a concerted campaign to spread doubt about the results of climate science as a dispute-independent debunking explanation of a belief that global warming is not real (Oreskes & Conway, 2010). But the mere fact that disagreements are clustered is not a good reason to downgrade someone's epistemic position.

At the same time, the objection from 'spinelessness' itself might seem to be somewhat puzzling. Instead of the negatively charged term 'spinelessness', we might as well construe conciliatory views as promoting intellectual humility, as Matheson (2015) suggests. The worry as it is stated is that it is implausible that epistemic rationality, in the broadly evidentialist sense usually at issue in the epistemology of disagreement, requires us to substantially reduce confidence in our political, philosophical, or moral views (Elga, 2007). But it is seldom argued at any length *why* it is implausible that epistemic rationality might require us to reduce confidence in this way. Apart from simply stating that a rational requirement of widespread reductions of confidence would be absurd, suggestions have been made that it is implausible because it would constitute an abdication of our epistemic responsibility, or be contrary to our integrity as believers (Aikin, Harbour, Neufeld, & Talisse, 2010; Pettit, 2006). But, on the theoretical resources available to us in the evidentialist framework, our epistemic responsibilities and our integrity as believers are plausibly in large part a matter of believing in line with our evidence. I tend to agree with Feldman (2006), Christensen (2014b), and Matheson (2015), that rationality on the evidentialist framework very well might really require us to be less confident about many controversial issues. While this is certainly an uncomfortable consequence, it's at least unclear to me how such discomfort could figure into an epistemic assessment on this framework.

In the final analysis, I do not think that widespread disagreement means that we are rationally required to be much less confident in our political beliefs. For one thing, we may have moral or prudential reasons for our doxastic attitudes, even if they should fall short of being supported by our evidence (Brogaard, 2014; Kahan, 2017; Lessig, 1995). Second, and closer to our present discussion, epistemic rationality itself does not, I think, require a very drastic decrease in our confidence in our political beliefs, because we should not be strongly committed to the evidentialist framework itself. As argued in article 3, conciliating in cases of disagreement can carry *epistemic* costs (including opportunity costs) to the agents – it can, for example, preclude their arriving in the near future at the doxastic attitude that is best supported by the

evidence. On notions of epistemic rationality that allow for a moderate form of epistemic teleology to figure in epistemic assessment, epistemic rationality does not recommend dramatically reducing confidence about our political and other controversial views in response to evidence of disagreement; at least not always.

3.2.2 Steadfast views

Another way to avoid the conclusion that we ought to abandon many of our political and moral views is to show that, even on the evidentialist framework, evidence of peer disagreement does not usually require us to conciliate. This is what defenders of steadfast views try to do. On steadfast views, it is not the case that both parties should reduce confidence in an idealized disagreement. What steadfast views have in common is the notion that there is always some breaker of the symmetry in the parties' epistemic position, allowing at least one person to give more weight to their own view than they do to their peer's. Like was the case for conciliatory views, there are many different motivations for steadfast views.

Kelly (2005) presents what has since become known as the Right Reasons View.[4] He argues that the higher-order evidence you get from disagreement does not defeat the justification for your doxastic attitude toward the disputed proposition, so you are not rationally obliged to reduce confidence. Suppose that you believe p on the basis of E. When you discover disagreement, you do not, on Kelly's view, get first-order evidence that p is false. You only get higher-order evidence that E does not support p. But the justification for your belief that p depends only on whether E supports p, not on whether your belief that E supports p is justified.[5] Of course, if E does not in fact support p, you should reduce confidence (indeed, you should defer completely to your interlocutor if E supports his or her belief), but this is not because of the evidence of disagreement, but simply because you should believe what your first-order evidence really supports. Kelly offers a number of arguments for the conclusion that the higher-order evidence constituted by peer disagreement does not give you evidence against p. One is that we do not typically cite higher-order evidence when giving reasons for our views. When I lay out my reasons for a belief that p, I refer to E. I do not include among my reasons the fact that I have inferred p from E. Another is that to countenance such higher-order evidence as being evidence

[4] The name is courtesy of Elga (2007).

[5] If your higher order justification is defeated but you are still justified in your belief that p, this would entail that you are justified in epistemic akrasia. A discussion of akrasia would take us too far afield, but see (Lasonen-Aarnio, 2014; Sliwa & Horowitz, 2015) for discussion.

for p would amount to double-counting the first order evidence. E is already integrated in my belief that p. To become more certain that p on the additional basis that I have inferred p from E would be to include E in my assessment twice.

A different argument for a steadfast view that Kelly (2005) presents does not rely on a denial that higher-order evidence can affect one's justification for believing p. Rather, the claim is that in a case of disagreement, the higher-order evidence that you have inferred not-p from E cancels out with the higher-order evidence that I have inferred p from E. Since the higher-order evidence cancels out, it does not give me a reason to reduce confidence in p. More specifically, the argument supposes that, prior to the discovery of disagreement, you believe p on the basis of E. When you discover disagreement, two more pieces of evidence enter the equation: The fact that I have inferred p from E, and the fact that you have inferred not-p from E. Kelly's claim is that the additional evidence gained by the discovery of disagreement does not warrant a change in doxastic attitude toward p. Giving equal weight to the higher-order evidence about you and the higher-order evidence about me means that they cancel out, and we are left with E as the basis for the belief that p, as we were before.

This argument is vulnerable to the objection that you have the higher-order evidence about yourself prior to the discovery of disagreement. If your doxastic attitude toward p already reflects this higher-order evidence about yourself, then the discovery of disagreement will change what you are justified in believing about p. And if we suppose that you only get the higher-order evidence about yourself when you discover disagreement, then the fact that you gain this piece of higher order evidence about yourself, not the hypothesis that disagreement has no epistemic significance, might explain why no doxastic change is required (Matheson, 2015).

Steadfast views have also been defended by appeal to the significance of the first-person perspective. Foley (2001), for example, suggests that self-trust entitles one downgrade one's estimate of the epistemic position of an epistemic peer when one discovers disagreement (this seems to amount to a denial of INDEPENDENCE). Wedgwood (2007) argues in a similar vein that the symmetry in a case of peer disagreement is illusory. From the first-person perspective, one is entitled to a degree of trust in one's own faculties that one need not give to the other. A slightly different, but related, defense of steadfast views comes from Plantinga (2000). When I am a party to a peer disagreement, I may realize that the evidence no longer supports p over not-p. Nevertheless, p still continues to seem to be true to me. Since we are fallible epistemic creatures and may err whichever doxastic attitude we end up adopting, we can do no better than believing what seems true to us.

As the discussion of Fumerton's (2010) debunking argument might suggest, I am not enthused by these arguments. Trusting one self more than one's interlocutor with respect to the disputed proposition does not seem warranted to me. In a case of peer disagreement, it is a salient possibility that I have made an error, and I have no reason to suspect that the error lies with you rather than me. For me to reject the possibility that I have erred because I am entitled to trust myself strikes me as dogmatic. Similarly, the fact that p continues to seem true to me is not a good reason to continue to believe p in a case of peer disagreement. It might be the case that I am ordinarily justified to presume that when something seems true to me, this gives me a reason to think that it is true. But this presumption is defeated by disagreement. For presumably it also continues to seem to you that p is false after we discover our disagreement. There is no reason then to suppose that my seeming is more accurate than yours.

A final line of argument for steadfastness that I will mention comes from the notion of reasonable disagreement and a denial of the so-called uniqueness thesis (Feldman, 2006; White, 2005):

UNIQUENESS: For a proposition p and a body of evidence E, there is at most one rational doxastic attitude toward p on E.

If UNIQUENESS is false, then it can be rational to adopt different doxastic attitudes toward p on the basis of E. So when I get evidence of a disagreement, it need not be evidence that either of us has responded to the evidence in a substandard way. If it is not, then there does not seem to be a reason for us to adjust our views (Rosen, 2001).

I will not engage in discussion of the truth or falsity of UNIQUENESS here.[6] It should be noted, however, that the mere falsity of UNIQUENESS would not push very far in the direction of steadfastness. If an extreme kind of permissiveness was true, and E always allowed any doxastic attitude toward p, then disagreement could never provide evidence that one's belief is unjustified. But such an extreme permissivism is not very plausible, and a more moderate permissiveness would not have this conclusion. There would always be *some* possibility that one of us has made a mistake. Suppose, for example, that for a proposition p the evidence E is permissive in the sense that any credence in the interval [.5, .9] would be rational. There is of course a risk that I make a mistake and arrive at a credence outside this interval, and in a case of

[6] For such discussion, I refer to (Kelly & White, 2014; Titelbaum & Kopec, 2017; White, 2005).

peer disagreement, this risk is the same for me as it is for you. Evidence of disagreement thus makes salient a possibility of error, which it seems that I should be responsive to.

3.2.3 Intermediate views

Some recent prominent views are hard to classify as either conciliatory or steadfast, although I am inclined to place them on the steadfast side of the center of the spectrum. I will mention the Total Evidence View of Kelly (2010, 2013),[7] and the Justificationist View of Lackey (2008b).

Recall that Kelly objects to purely conciliatory views on the ground that they demand that first-order evidence is 'swamped' by higher order evidence in determining what doxastic attitude is rational. His Total Evidence View seeks to restore first-order evidence to epistemic significance in cases of disagreement. Kelly argues that in a case of peer disagreement, the rational doxastic attitude is a function of both the higher-order evidence and how well one initially responded to the first-order evidence. If you initially responded correctly to E, then you should be less moved by the disagreement than I should if I responded incorrectly. You might not be in a position to ascertain that you have responded correctly or incorrectly, but rationality does not always supervene on phenomenological states. Defenders of conciliatory views might object that this is unhelpful. When you are a party to a peer disagreement, you have no way of knowing whether you or your peer responded correctly to E. If so, how could you rationally be less moved than conciliatory views would have it simply because you did respond properly? Another worry is that, although there might be an asymmetry in justification at the first-order level, symmetry is restored at the higher-order level. Neither party's belief that they have responded properly to the evidence is more justified than the other party's. Given this, it would not be reasonable to give less than equal weight to your peer. Kelly rejects this line of reasoning. He argues that the degree of justification one has in one's first order belief (how well one has initially responded to E), influences how justified one is in a higher order belief that one has responded properly to the E. If I have responded properly to E and you have not, then my belief that I have properly responded to E is more justified than your belief that you have properly responded to E. How justified one is in such a higher-order belief is therefore not just a matter of track-record, familiarity with the evidence, or intellectual virtue, as conciliatory views would typically have it, and so symmetry is not restored at the higher-order level (Kelly, 2010).

[7] Kelly has abandoned his earlier, more steadfast, Right Reasons View (Kelly, 2005) mentioned in the above section.

Kappel (2017) has recently argued against this latter notion of "upward epistemic push". According to Kappel, it is not the case that justification at the first-order level affects justification at the higher order level. In contrast, justification at the higher-order level can affect justification at the first-order level, such that if your justification for a higher order belief that you have properly responded to E is defeated (e.g. by evidence of disagreement), then your justification for believing p is also defeated, even if it were initially stronger than your interlocutor's.

Lackey's (2008b) Justificationist View bears some resemblance to the Total Evidence View. Her view has it that the epistemic significance of disagreement depends on the degree of justified confidence with which a view is held. In addition, it holds that personal information, things that you know about yourself but do not know about your interlocutor, can sometimes act as a symmetry breaker. On the Justificationist View, no doxastic revision is required if and only if your belief that p has a high degree of justified confidence, and you have a relevant symmetry breaker. In contrast, substantial doxastic revision is required if your belief that p has a low degree of justified confidence. A moderate amount of doxastic revision is required in cases that straddle the middle of this interval. The notion of justification at play here is not entirely internalist. Reliability or truth-conduciveness of the belief-forming mechanism responsible for your belief is a requirement for justification. So when two epistemic peers disagree, differences in the justification of their belief can make a difference to how they ought to respond, if one of them has a 'symmetry breaker' available.

Lackey draws on the different verdict cases like RESTAURANT yields compared to more extreme cases like the following:

EXTREME RESTAURANT: While dining with four of my friends, we all agree to leave a 20% tip and to evenly split the cost of the bill. My friend, Mia, and I rightly regard one another as peers where calculations are concerned—we frequently dine together and consistently arrive at the same figure when dividing up the amount owed. After the bill arrives and we each have a clear look at it, I assert with confidence that I have carefully calculated in my head that we each owe $43. In response, Mia asserts with the same degree of confidence that she has carefully calculated in her head that we each owe $450, which is more than the total cost of the bill (2008b, p. 321).

The Justificationist view has it that in this case, my belief that the share is $43 enjoys a high degree of justified confidence, due to my track record of successfully calculating shares. In addition, I have personal information that can act as a relevant symmetry breaker. I know that I am being sincere, and that I am not intoxicated or otherwise incapacitated at the moment. But in light of your rather extreme response, I have no such knowledge of you. For all I know, maybe you are being insincere, or maybe you went to the bathroom during lunch and took hallucinogenic drugs. This asymmetry in personal information, combined with my high degree of justified confidence, means that no doxastic revision is required in this case. In contrast, in RESTAURANT, while my belief that the share is $43 enjoys an equally high degree of justified confidence, I have no symmetry breaker available. I know that I am not intoxicated or incapacitated, but I have no reason to suspect that you are either. So I am required to revise my doxastic attitude.

3.3 Evidence and epistemic teleology

This concludes my very selective overview of the epistemology of disagreement. Articles 2 and 3 both tackle the core question. As mentioned, article 2 addresses a problem that arises about how to determine the epistemic significance of disagreement about politically divisive propositions due to the influence of motivated reasoning on our beliefs about politically charged facts. Because of motivated reasoning, we should think that the familiarity with the evidence and intellectual ability of someone from the other side of the political aisle is inversely correlated to their probability of being right. Thus, determining the extent of doxastic revision required upon the discovery of such disagreements requires a choice between one of these notions, but either option has problematic implications.

Where article 2 discusses a problem about the epistemic significance of disagreement within the terms set by that debate as it is traditionally construed, article 3 breaks with those terms to some extent. It also attempts to provide an answer the core question, specifically for cases where the discovery of disagreement is followed by deliberation with one's interlocutor(s). But its answer does not depend on any particular conclusion about what our evidence supports in a case of disagreement. Instead, it presents a novel argument for the rationality of maintaining confidence in response to evidence of disagreement based on a version of epistemic teleology. Both parties' maintaining confidence in such cases promotes one's arriving at the doxastic attitude that is best supported by the evidence. It does so through facilitating a division of epistemic labor in collective deliberation that improves the group's ability to find and evaluate

evidence. The fact that maintaining confidence promotes one's belief having the epistemic value of being supported by one's evidence makes maintaining confidence epistemically rational. In terms of substantive conclusion, article 3 thus expounds a steadfast view. However, its route to this conclusion is, in a certain sense, orthogonal to the standard debate between proponents of conciliatory and steadfast views. The standard debate is, as I have construed it, about what doxastic revision one's evidence, first-order and higher-order, supports in cases of disagreement. Article 3 takes no stance with respect to that question – it is as compatible with a conciliatory response as it is with a steadfast one. Indeed, in spite of article 3 having a substantively steadfast conclusion about what doxastic revision is required, I am inclined to favor a conciliatory response to the question of what one's evidence supports in cases of disagreement. Article 3 argues that, whatever the correct verdict in the standard debate may be, there are reasons of an epistemic teleological kind that make it all things considered epistemically rational to maintain confidence, even if the evidence of disagreement, taken in isolation, does not support maintaining confidence. So it may be that in a case of disagreement, one's evidence supports conciliating. But, article 3 argues, what one's evidence supports does not settle the question of what one ought to believe. Indeed, it argues that evidential concerns can be overridden by other concerns in determining the epistemic rationality of doxastic attitudes.

An argument put forward in article 1 addresses a question that is somewhat related to the core question, but has received nowhere near the same amount of attention in the philosophical literature: Namely what (if any) effect evidence that one is party to a disagreement (in the weak sense that one is aware that one's belief is controversial) has on the normative evaluation of one's biased reasoning about subsequently encountered evidence. It responds to an argument by Kelly (2008) to the effect that when we are justified in believing p, we are justified in suspecting that evidence against p is flawed, so we are justified in selectively scrutinizing this evidence. Article 3 argues that, while one may be rational in selectively scrutinizing evidence against a belief whose truth one can justifiably take for granted, disagreement precludes any such justification. So, in cases where one evaluates evidence pertinent to a controversial belief, one is not rational in selectively scrutinizing evidence according to whether it counts in favor of one's belief. Thus, awareness of disagreement (in the weak sense) can have defeating force with respect to what kinds of reasoning it is rationally permissible to engage in. Now, an interesting point that is not discussed directly in the articles themselves is that, for the reasons presented in article 3, there are circumstances, namely those involving (anticipated) collective deliberation, where the kinds of reasoning, and their resulting doxastic attitudes, judged to be irrational by the

standards of article 1 *are* rational when we take more teleological standards into account. Together, articles 1 and 3 suggest that the social circumstances we find ourselves in can make a critical difference for our evaluations of epistemic rationality. I return to this issue in the concluding remarks.

4 Disagreement and democracy

In addition to the debate in epistemology, disagreement has played a prominent role in political philosophical discussions. Article 4 addresses one of these discussions, namely the one about the proper role of our controversial commitments and beliefs in liberal democratic decision-making. A few words are perhaps in order about this change in focus, and about the connections between the epistemic and the political topics. In addition to a common thread of disagreement and motivated reasoning, there is, as I see it, a relatively straight line connecting more specific issues dealt with in articles 1 and 2, and 3, respectively, to those of article 4. All of articles 1, 2, and 4, pertain to motivated reasoning about politically charged facts. One topic of article 4 is the potential of 'deliberative debiasing' to diffuse factual political disputes, the discussion of which relies on some of the same findings about collective rationality in disagreeing groups that feature prominently in article 3.

The topic of article 1 and, albeit indirectly, article 2, is belief polarization: the tendency for disagreements to grow more extreme due to confirmation bias and motivated reasoning in our treatment of evidence. As mentioned in section 2.3 of this introduction, belief polarization is typically observed, both within the psychological laboratory and in society at large, for topics that divide us along political and cultural fault lines. Article 4 addresses what the presence of sharp, politically structured disagreement about these issues means for democracies faced with having to make decisions that rely on a substantive view about the disputed facts. In particular, what decisions are politically legitimate in democracies when there is a consensus among experts that affirms the factual beliefs of one side of the political or cultural divide? Article 3 relies for its conclusion on the epistemic benefits of disagreement to the results of collective deliberation. Doxastically diverse groups do much better than homogeneous groups when they engage in collective reasoning. Where article 3 used this finding to motivate a response to the question about the epistemic significance of disagreement, article 4 uses it to discuss the potential of 'deliberative debiasing' for diminishing widespread factual disagreements and help citizens arrive at factual beliefs that are in line with our best available science.

Thus, while the articles might at first glance seem to fall into two sharply divided groups, I think there are relevant connections across the political-epistemic divide. In what follows, I will provide a very brief presentation of an issue raised by the presence of disagreement in democracies: that pertaining to political legitimacy and the kinds of justifications that can be offered for policies about which we disagree. This presentation will be much briefer than that of the epistemology of disagreement. This is both because only one of the four articles pertains to

this issue, and because the core political-philosophical discussion in article 4 is due my co-author on that article, Andreas Christiansen.

4.1 Disagreement and public reason

Disagreement is a fact of life in the political domain. We disagree about matters such as the basic values that ought to guide our institutions and policies, about whether specific policies are morally right or wrong, and about factual questions that are perceived as relevant to policy.

Disagreement in the public sphere is often taken to pose a problem to political legitimacy. Political legitimacy is, at least on one view, a question of when political authority or coercive power is justified and imparts obligations on the governed (Rawls, 1993). Political legitimacy can be interpreted as a descriptive concept, i.e. as a question of what features of institutions and policies lead to their being considered legitimate by the public (Weber, 1964). While article 4 will touch upon this, it is mainly concerned with a normative conception of political legitimacy: the question of what features institutions and policies are required to have for them to be justified in exercising political authority or coercive power (Rawls, 1993). While it is relatively clear that a policy would be legitimate if the entire population agreed that it was right, disagreement raises the thornier question of when it is legitimate to exercise power over citizens who disagree with a policy. It seems that, in order for policy to be legitimate, it should, at least in principle, be justifiable to all those it would exercise power over. Without this requirement majorities could, for example, institute grossly unjust policies at a minority's expense without offering them reasons that they would recognize as speaking in favor of the policy.

One solution to this problem that has been proposed is the ideal of public reason. Roughly, public reason is the idea that the justifications for an institution or policy must be grounded in reasons or arguments that all citizens (at some level of idealization) would recognize as speaking in favor of the policy (Gaus, 1996; Rawls, 1993). Public reason can be construed as both demanding that policies actually be justifiable in this manner, but also as demanding that the justifications that the authorities offer for them in public discourse be restricted to this kind, or that citizens restrict themselves to public reasons when they offer arguments for policies to other citizens.

So what qualifies as public reasons in this sense? It is perhaps easier to answer this question by referring to what does not qualify. On Rawls' (1993) view, what does not qualify as public reasons are those based on controversial moral, religious, or political doctrines that

reasonable (permissively and broadly construed) people disagree about.[8] A reason for a policy of putting apostates to the sword based on it being the will of God that apostates be put to the sword, according to controversial doctrine D, would not be a public reason. When it comes to questions of value, public reasons are those that refer to values that all reasonable people can endorse. This is taken to include basic values such as freedom, equality, and avoidance of unnecessary harms (Quong, 2018), but to exclude things like the value of obeying the will of God and the values of Nazism. When it comes to facts, public reason includes facts produced by science. So a reason for a policy of restricting halocarbon refrigerants based on their contribution to ozone layer depletion is a public reason, as it based on our best science and is one that all reasonable people would see as speaking in favor of the regulation of such chemicals. A complex and largely unsettled question is whether scientific facts that are controversial are part of public reason. Take the proposition that the death penalty is a deterrent to murder. This is not based on any controversial moral, religious, or political doctrine, but there is reasonable disagreement, even among scientific experts, about whether it is indeed a fact (National Research Council, 2012). On a standard that excludes controversial scientific claims from public reason (Rawls, 1993), such a proposition might not be part of public reason, although it is unclear exactly what kinds of controversy about scientific facts can exclude them from this domain (Jønch-Clausen & Kappel, 2016). For example, the reality of anthropogenic global warming is a controversial scientific fact in the sense that a substantial proportion of the population disagrees with it, although it is not controversial among the scientific experts themselves. Does this public controversy exclude the fact that anthropogenic climate change is occurring from public reason?

This is one of the primary questions addressed in article 4. In particular, we discuss what impact it has on the legitimacy of policies that they rely for their justification on scientific facts that, while uncontroversial among experts, are controversial among the public *because* they are entangled in a dispute between controversial cultural outlooks. Does this entanglement mean that the factual beliefs, which are in a certain sense *caused by* or *expressive of* the controversial outlooks that are typically excluded from public reason, themselves should be excluded, even when those factual beliefs are reflective of the best scientific evidence? A related question is to what extent policy decisions in a democracy should be responsive to factual beliefs when they are in conflict with scientific consensus and we know roughly why: because they are the result of people's cultural commitments. Is it a requirement for a state to be a democracy that it is

[8] Gaus (1996) has a more permissive notion of public reason that would include these.

responsive in the sense that policy reflects these (false) beliefs, or can elected governing bodies legitimately ignore them and not lose their status as democratic?

5 Reflections on psychological research in philosophical argument

Both the epistemological and political-philosophical discussions of disagreement have heretofore largely been carried out in a highly idealized manner that attempts to abstract away as many gritty and potentially contaminating details as possible for the purpose of achieving clarity. My PhD project is part of a larger project, spearheaded by Klemens Kappel and funded by the Danish Free Research Councils, that aims to broaden out the epistemological debate to include cases of 'complex disagreement', where the usual idealizations fall short of obtaining in various ways. In particular, my role in this project has been to supply knowledge of relevant psychological findings and explore their implications for the philosophical debate.

I quickly discovered that this is was not an unproblematic endeavor. For example, one of my sub-projects in the initial project application was to investigate how people actually respond to evidence of disagreement of the type discussed by epistemologists. Two problems soon presented themselves, however.

The first was that it is not a question that psychologists have given very much attention. This is perhaps partly a result of a lack of interest in the topic, but it is also something of a methodological nightmare to study this question. It is extremely difficult to experimentally control for variations in epistemic position, and to track subjects' doxastic states and their causes over time in the way that would be required to get results that would be of genuine interest to the epistemological discussion. The studies on conformity in the tradition of Asch (1956) are those that to my knowledge come closest to investigating responses to disagreement in a way that is somewhat similar to the cases discussed by epistemologists. But while some people in such experiments tend to act in ways that could be construed as 'conciliatory' (and others don't), it is extremely difficult to know the degree of conformity that is attributable to genuine belief change rather than something else. And even if conformity were purely a reflection of genuine belief change, and we could somehow know this, it would be hard to know the extent to which this belief change is caused by the kinds of considerations that epistemologists take to be relevant, rather than things like perceived social pressures or a desire to not stand out from one's group (Abrams et al., 1990).

The second was that, even if there were many relevant studies about how people do respond in the relevant kinds of case to be found, it is questionable what normative significance

this would have. Perhaps proponents of conciliatory (or steadfast) views would interpret a finding that people do typically respond to disagreement in the way that their view suggests that they should as 'evidence' in its favor. But proponents of the competing view could simply respond that it only shows that people are regrettably irrational in those kinds of cases. While such findings might be interesting in their own right, the aim of the project was not to merely describe people's reactions to disagreement for its own sake, nor was it to 'naturalize' the epistemology of disagreement in the sense of Quine (1969).

So instead of looking at how people really do respond in the kinds of cases that have attracted philosophical interest, I decided to look at whether there might be other types of psychological research that could have implications for the normative arguments and conclusions in the traditional debate. In this endeavor, which has ultimately resulted in the contents of this dissertation, I also encountered early problems, mostly stemming from my having an educational background in psychology and therefore being somewhat underprepared for the degree of rigor required of philosophical argument. For example, I was initially inclined to think that it was a good argument for a steadfast response to disagreement that I just point at the psychological studies demonstrating the positive consequences of disagreement for group deliberation and leave it at that. When it was pointed out to me that such consequences are typically not part of epistemic assessment, and I would have to show why they ought to be, it was initially frustrating but ultimately illuminating. I hope that I have ended up with a body of work that demonstrates that psychology can be relevant to philosophy on philosophy's own terms, and that it can do more than just complicate matters: It can fundamentally alter our normative verdicts, sometimes in surprising ways.

Of course, as is always the case with empirical findings and any philosophical arguments that rely on them, there is a possibility that the findings that I drawn on will turn out to be wrong. This possibility is particularly salient in light of the ongoing replication crisis in psychology and science at large (Ioannidis, 2012; Klein et al., 2014; Open Science Collaboration, 2015). To the best of my knowledge, none of the findings I have drawn on are cast into doubt by failed replications, but they might be tomorrow. Until they are, I think that the epistemology and political philosophy of disagreement do well to take them seriously.

I'll close this introduction by considering the question of whether my own motivated reasoning has shaped the arguments in the dissertation. The short answer is: Of course! It is impossible (at least for myself) not to become somewhat committed to philosophical positions you take up defense of, and this commitment is bound to skew any subsequent reasoning about the matter in

some way. Nor do I have any illusions that my awareness of the various biasing factors somehow inoculates me from their influence. The extent to which this has caused me to make mistakes that invalidate the conclusions is up to the readers to ascertain. I can only hope that some readers are inclined to disagree with the conclusions and so are inclined to work extra hard to find the flaws that are doubtlessly there. That could initiate a process of collective deliberation that would hopefully ultimately make both parties wiser.

Article 1: Belief polarization and congeniality bias in reasoning

1 Introduction

The beliefs of individuals who disagree sometimes polarize after the individuals are exposed to the same body of evidence.[9] Experiments by psychologists, political scientists, and economists, have repeatedly demonstrated this belief polarization phenomenon for many different domains of belief and different types of evidence.[10] Outside of the lab, beliefs about many important matters of fact are polarized among the general public, and such disagreements often persist or widen even as public evidence relevant to the matter at hand accumulates. For example, beliefs among the US public about the reality of anthropogenic climate change have polarized along political lines since the 1990s, all the while there has been a dramatic increase in the amount of publically available relevant research (McCright et al., 2014; Pew Research Center, 2016).

While most work on belief polarization is purely descriptive, its normative status has attracted commentary since some of the earliest studies demonstrating the phenomenon. For instance, Ross & Anderson (1982, p. 145) claim that belief polarization is "in contrast to any normative strategy imaginable for incorporating new evidence relevant to one's beliefs." While less strong in their condemnation, Lord et al. (1979, p. 1107) hold that their subjects "sinned" in "their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially 'justified' the processing bias." Indeed, at a first glance, belief polarization might seem to be a sign of obvious irrationality. One of the roles often attributed to evidence is to act as a neutral arbiter in disagreements (Kelly, 2016). When agents base their beliefs on evidence, we expect disagreements to be fragile: as more shared evidence emerges, such agents ought to converge toward the view best supported by the total evidence. Belief polarization therefore seems to suggest that the agents do not properly base their beliefs on evidence.

However, subsequent treatments of the normativity of belief polarization have largely emphasized that belief polarization can be epistemically rational, both in theory and in the kinds

[9] I mean "disagree" in the weak sense that the individuals have different doxastic attitudes toward some proposition. It is not necessary that the individuals are aware of this dispute, or even each other's existence.

[10] For illustrative findings, see (Andreoni & Mylovanov, 2012; Batson, 1975; Cook & Lewandowsky, 2016; Jern et al., 2014; Kuhn & Lao, 1996; Liberman & Chaiken, 1992; Lord et al., 1979; Miller et al., 1993; Munro & Ditto, 1997; Munro et al., 2002; Plous, 1991; Pomerantz et al., 1995; Taber & Lodge, 2006).

of actual cases studied experimentally. One argumentative route to this conclusion has been to contend that selective scrutiny of evidence uncongenial to one's prior belief, rather than being an irrational bias, is quite rational in light of one's expectations about the quality of the evidence. Such scrutiny can give rise to alternative explanations of the evidence that decrease its impact on the target proposition. Meanwhile, supporting evidence is given full weight, as the lack of scrutiny means that no alternative explanations for this evidence are uncovered. Since alternative explanations that one uncovers are themselves part of one's total evidence, polarized beliefs that arise on this basis can be a rational response to one's total evidence (Kelly, 2008). Another line of reasoning shows that biased evaluations of evidence and belief polarization can be consistent with Bayesian updating, and that several experimental results can be accounted for in this manner (Andreoni & Mylovanov, 2012; Baliga, Hanany, & Klibanoff, 2013; Cook & Lewandowsky, 2016; Jern et al., 2014; Olsson, 2017).

This article presents reasons to think that belief polarization, and the biased evaluations of evidence that are its cause, are more problematic than these authors suggest. Kelly's (2008) arguments for the rationality of selective scrutiny and the resulting belief polarization either fall short of showing that they are rational, or fail to apply in the standard cases of belief polarization that have attracted interest. And while it may be possible to model canonical cases of belief polarization as being consistent with Bayesian updating, the set of priors that allow agents in such models to reproduce the results are themselves often highly problematic. A further problem for both lines of argument is that they rest on a construal of the psychological underpinnings of belief polarization that emphasizes prior belief as the sole cause of biased evaluations of evidence. The general applicability is of this model is disputed by a broad array of psychological research, which shows that non-epistemic motivations and emotion typically play much larger roles in explaining belief polarization. While it may be the case that rational belief polarization is not an oxymoron, the kinds of cases that have been demonstrated experimentally, and those that figure prominently in the public sphere, are not instances of such.

2 The prior belief model of belief polarization

The most widely known demonstration of belief polarization, and the one that has formed the basis for much of the subsequent normative discussion, was conducted by Lord et al. (1979). Participants had been pre-tested to identify proponents of the death penalty who believed it was deterrent to crime, and opponents of the death penalty who believed that it was not. All participants were asked to assess a set of mixed evidence: two studies presenting statistical data

on the efficacy of the death penalty as a deterrent to murder. One study supported a deterrent effect, while the other did not. While the methods of the studies were slightly different, they were roughly equal in quality, and the researchers varied which type of study supported which conclusion between participants. Lord et al. found evidence of biased evaluation of evidence: After assessing both studies, most participants found the study that confirmed their prior belief to be methodologically superior to, and more persuasive than, the study that disconfirmed their prior belief, both as measured on a numerical scale and by listing thoughts about the studies.[11] Later studies have corroborated that this difference is due to selective scrutiny: subjects tend to spend much more time evaluating evidence against their prior belief than evidence in its favor, and predominantly spend their time disparaging the quality of the disconfirming evidence and bolstering the evidence in their favor (Taber & Lodge, 2006). In both Lord et al.'s study and many that followed, the result was belief polarization: Proponents and opponents reported having become more extreme in their beliefs.[12]

The most well-known philosophical discussion of this case comes from Thomas Kelly (2008). Kelly precedes his discussion of its normative implications with reflections on the psychological underpinnings of biased evaluation of evidence and belief polarization, emphasizing how prior belief can shape one's treatment of subsequently encountered evidence. Individuals who believe p will tend to believe that there are no sound arguments for not-p. Upon encountering an argument for not-p, such individuals will therefore be disposed to treat the argument with suspicion, and to expend cognitive resources in an attempt to uncover any flaws that show it to be unsound, whereas they will not be similarly disposed when encountering arguments for p. In the same vein, individuals who believe a hypothesis H will be inclined to suspect that there are alternative explanations of evidence for which not-H is a potential explanation, rather than immediately take not-H as the actual explanation. When such individuals encounter evidence for

[11]A similar phenomenon is belief bias in argument evaluation: Subjects are more likely to judge formal arguments whose conclusion agrees with their prior as being valid, and informal arguments whose conclusion agrees with their prior as being stronger (Evans et al., 1983; Klauer, Musch, & Naumer, 2000; Thompson & Evans, 2012).

[12] It is perhaps noteworthy that Lord et al. (1979), like several other studies, found evidence of belief polarization only for a measure of self-reported belief change, rather than by comparing measures of belief taken before and after assessment of the evidence. See (Gerber & Green, 1999; Miller et al., 1993; Ross, 2012) for discussion of the significance of this. Other studies do however find direct evidence of pre-post belief change (Batson, 1975; Pomerantz et al., 1995; Taber & Lodge, 2006).

which not-H is a potential explanation, they will therefore be disposed to scrutinize it in search of any alternative explanations of the evidence, for example flaws with the methods or analysis of a study.[13] Suppose that such scrutiny is successful in uncovering alternative explanations. Awareness of these competing alternative explanations then decreases the extent to which the evidence is taken by the individual to confirm not-H. In contrast, when such individuals encounter evidence for which H is a potential explanation, they are likely to conclude that H is the actual explanation of the evidence, without engaging in scrutiny. When encountering a body of mixed evidence similar to participants in Lord et al.'s study, such individuals are likely to bolster their belief in H, since they will tend to give more weight to the confirming evidence than the disconfirming evidence. In a similar manner, individuals who initially disbelieve H are likely to bolster their belief in not-H in response to the same body of evidence, yielding belief polarization.[14]

The adequacy of this psychological description is challenged in section 3.3. But for the time being, let us proceed under the assumption that it is an accurate reflection of subjects' cognitive processes in the relevant experiments. Kelly's subsequent normative discussion tackles two separate questions. The first is whether, on the described psychological picture, selectively scrutinizing evidence that is incongruent with one's prior belief is rational or not. The second is whether the beliefs that result from such selective scrutiny are rational or not. The first question

---

[13] Kelly quite correctly describes this as the "default state", and likely proceeding without conscious awareness, rather than as a deliberately adopted strategy.

[14] There are cases of belief polarization that result not from mixed evidence, but from a single piece of evidence. For example, Batson (1975) had subjects with varying degrees of Christian belief read a story about clergy conspiring to cover up evidence undermining the divinity of Jesus. After reading the story, subjects with low degrees of Christian belief expressed even lower degrees of belief, as expected, while strongly Christian subjects expressed even stronger religious beliefs. More recently, Nyhan, Reifler, Richey, & Freed (2014) found that parents least favorable toward the MMR vaccine became less likely to vaccinate a child in the future after receiving a message correcting misperceptions about vaccines being a cause of autism, while other parents became more likely to do so. Kelly does not discuss such cases, but a similar type of account can possibly be offered for the underlying psychology. When individuals who believe H encounter evidence for which not-H is a possible explanation, they are disposed to scrutinize the evidence in search of alternative explanations. Suppose that in the course of generating these alternative explanations, they also actively counter-argue, and so come upon reasons to believe that H is true, which they had not previously considered. If the support the evidence lends to not-H is sufficiently diluted by alternative explanations, and these novel reasons provide considerable support for H, such individuals might increase their credence in H after encountering evidence against H.

concerns whether a certain mode of reasoning is legitimate, and is, according to Kelly, primarily a question about practical rationality. The second concerns whether the beliefs so produced are properly based on one's evidence, a question of epistemic rationality. The present discussion will follow the same structure.

3 Biased evaluation of evidence

We begin with the first question: Whether engaging in selective scrutiny of evidence that is uncongenial to one's prior belief is rational in cases of belief polarization. Interestingly, Lord et al. (1979, p. 2106) seem to suggest an affirmative answer early in their brief discussion of the normative implications of their results: "[T]here can be no real quarrel with a willingness to infer that studies supporting one's theory-based expectations are more probative than, or methodologically superior to, studies that contradict one's expectations. When an 'objective truth' is known or strongly assumed, then studies whose outcomes reflect that truth may reasonably be given greater credence than studies whose outcomes fail to reflect that truth." For example, people are, they say, quite reasonable in being skeptical about reports of miraculous virgin births or herbal cures for cancer. An inclination to process new data in light of prior belief is "essential for any organism to make sense of, and respond adaptively to, its environment."

3.1 Selective scrutiny and scientific practice

Kelly's answer to the first question is likewise affirmative. He argues to this conclusion by an analogy with scientific practice. Scientists do not generally focus equally on all phenomena, but rather routinely engage in what appears to be selective scrutiny in their inquiry. This is because scientific practice is, at least to some extent, driven by anomalies. Phenomena that are not well accounted for by an accepted theory, or that on the face of things seem to amount to evidence against it, attract much more attention from scientists than do phenomena which are straightforwardly explained by existing theory. When scientists focus on anomalies, their efforts are directed at attempts to formulate and test hypotheses that would allow the anomalous phenomena to be brought into the purview of the accepted theory. Kelly's verdict is that scientists are rational when they proceed in this manner, even suggesting that any other practice would be irrational. And he asserts that the practice bears a striking similarity to the psychological picture of subjects in a belief polarization experiment, who selectively scrutinize the evidence that is not well accounted for by their prior belief or hypothesis in an attempt to

generate alternative explanations for such evidence. Given that scientists are rational in the way they proceed, and absent any reason to hold scientists and ordinary thinkers to different standards, ordinary thinkers are therefore rational in selectively scrutinizing evidence on the basis of how it accords with their prior belief.

Is it really the case that the scientific practice of focusing on anomalies is relevantly similar to selective scrutiny of the kind found in the case of belief polarization, to the extent that they warrant the same normative verdict? In a belief polarization experiment, we are assuming that selective scrutiny occurs due to a suspicion that a study for which the falsity of one's hypothesis is a potential explanation must contain some methodological flaws. The relevant suspicion here is that there is something wrong with the evidence itself – reasons to disregard it or at least give it less weight, and the subsequent scrutiny aims at uncovering these reasons. But this is not quite what Kelly (2008, p. 624, emphasis mine) describes scientists as doing when they engage with anomalies. Scientists attempt to "generate hypotheses that *allow the existence* of the anomalies to be reconciled with the currently accepted theory." The scientists' goal is not, or at least not exclusively, to deny the reality or pertinence of the anomalous phenomena by discovering flaws in the relevant studies. Rather, it is at least sometimes the case that the scientific work involves refining or amending the theory in order to accommodate the anomalous findings. If the sole purpose of scientists' directing their efforts at anomalous phenomena were to attempt to explain them away, then this would hardly be rational. But this is what we are assuming subjects typically do in a case of belief polarization.

There is another relevant difference between the selective scrutiny observed in experiments and scientific practice. In experiments such as (Lord et al., 1979), selective scrutiny occurs when subjects are exposed to a set of evidence that they are instructed to evaluate. Subjects do not choose whether to observe this set of evidence rather than some other set. However, scientists are not limited in this way in their practice. Their situation might, at least in part, be more accurately described as a choice of what evidence to attend to and what to ignore. Given this choice, scientists choose to attend to anomalies that seem to disconfirm the accepted theory rather than attend to various innocuous phenomena that are already explained by, and confirm, the accepted theory. To the extent that this description is accurate, scientists are proceeding in the *opposite* manner to ordinary individuals studied in similar choice situations. When given the choice between observing evidence that they expect to confirm their prior belief and evidence that they expect to disconfirm it, subjects in experiments overwhelmingly prefer to attend to confirming evidence. This is so even when they are explicitly instructed to (and presumably genuinely try

to) be even-handed, or when they are given monetary incentives to observe disconfirming evidence (Frimer et al., 2017; Hart et al., 2009; Taber & Lodge, 2006). Construing scientific practice in part as a matter of evidence selection and using it as a normative benchmark condemns, rather than vindicates, the kinds of thinking that lead to belief polarization.

A final, and perhaps more philosophically interesting, problem with the analogy is that it downplays the element of controversy at the level of theory at work in typical cases of belief polarization. In science, it is not always the case that there is a single currently accepted theory, on which some phenomena are innocuous whilst others are anomalous. Often, there are at least two seemingly plausible competing theories, which might pick out different phenomena as being anomalous. The question of whether the death penalty is a deterrent to murder is perhaps itself an example of this, as it remains unsettled among experts (National Research Council, 2012). When the task at hand for a scientist is adjudicating between two plausible competing theories, it would not be a rational way to proceed to take findings that appear congenial to one's favored theory, but are at odds with the competitor, at face value, while scrutinizing findings that appear contrary to one's favored theory but could be explained by the competitor.[15] The situation of subjects in belief polarization experiments typically resembles this situation in that the evidence pertains to beliefs about issues that subjects know to be at least somewhat controversial, such as stereotypes about homosexuality, the death penalty, gun control, or anthropogenic global warming (Cook & Lewandowsky, 2016; Hart et al., 2009; Kahan et al., 2012; Kahan, Peters, et al., 2017; Lord et al., 1979; Munro & Ditto, 1997). Thus, even if scientists were rational in expending more cognitive resources on engaging with anomalies in cases where there is one commonly accepted theory, this is seldom the situation facing subjects in cases of belief polarization.[16]

The element of controversy is also the reason why Lord et al. (1979, p. 2106) do not ultimately condone their subjects' biased evaluation of the evidence. While they countenance

---

[15] At least this seems to be the case when looking only at individual scientists. Selective scrutiny may be a boon to the scientific community as a whole insofar as it promotes a productive division of cognitive labor (Kitcher, 1990; Muldoon, 2013).

[16] There are exceptions. For example, some studies of belief polarization (e.g. Jern et al., 2014) ask subjects to evaluate evidence on a question that is completely novel to them prior to the experiment. Prior beliefs are manipulated by giving subjects different data in advance of their evaluation of evidence. Being unaware that other subjects are given different data, these subjects might rationally believe that the hypothesis supported by their initial information is the only plausible candidate, and assuming that their prior beliefs were rational responses to the initial data, these might be cases of rational belief polarization. They are, however, quite different from the typical cases and from belief polarization as observed in society at large.

biased evaluation when an "objective fact is known or strongly assumed", they find it doubtful that their own subjects reasonably fulfilled this condition, in light of the inconclusive nature of existing data and the widespread disagreement.

The analogy from scientific practice does not, in sum, constitute a very strong argument for biased evaluation of evidence being rational in standard cases of belief polarization. It may be the case that we are rational in selectively scrutinizing evidence against beliefs whose truth we can justifiably take for granted, such as the impossibility of virgin birth. But, with a few exceptions, it will not generally be true that two parties in a case of belief polarization can justifiably take the truth of their opposing views for granted in the relevant sense.[17]

3.2 Evaluation of evidence: Bayesian, not biased?

In contrast to Kelly, and most psychologists, scholars who have discussed belief polarization from a Bayesian framework do not rely on the notion of selective scrutiny of evidence based on its incongruence with prior belief. For these scholars, biased evaluation of evidence, in the purely descriptive sense that evidence consistent with one's prior beliefs tends to be considered stronger and more probative than evidence against one's prior belief, can simply be a consequence of standard probabilistic inference (Andreoni & Mylovanov, 2012; Cook & Lewandowsky, 2016; Gerber & Green, 1999; Jern et al., 2014; Olsson, 2017).

---

[17] Kelly appears to agree that we cannot often take our beliefs for granted when they are challenged in his discussion of Kripkean dogmatism as a potential explanation of belief polarization. Kripkean dogmatism is roughly the following puzzle: A person who has a justified belief that p at time t_0 would, on a closure principle about justification, also be justified in a belief that subsequently encountered evidence for not-p is misleading. The Kripkean dogmatist could therefore justifiably reject any such evidence encountered later. Kelly rejects Kripkean dogmatism as an explanation of belief polarization in favor of the selective scrutiny model, but notes that if belief polarization were the result of Kripkean dogmatism, it would be irrational. Kelly says, following Gilbert Harman, that Kripkean dogmatism is irrational because once the individual actually receives evidence against p at time t_1, the justification for the belief that p is defeated, and therefore so is the justification for the belief that evidence against p must be misleading. Kelly does not extend this line of reasoning to his discussion of selective scrutiny, but it seems that it is equally pertinent here. A person who has a justified belief that p at t_0 would, by closure, be justified in the belief that there must be alternative explanations for any subsequently encountered evidence for not-p, making it rational to selectively scrutinize such evidence. However, upon encountering evidence for not-p at t_1, the justification for the belief that p is defeated, and along with it the justification for the belief that there must be some alternative explanations that would account for evidence for not-p. To the extent that selective scrutiny relies on such a justified belief to be rational, selective scrutiny is no more rational than Kripkean dogmatism.

For a Bayesian agent, the direction of belief change in a binary hypothesis H warranted by a new piece of evidence E is determined by the likelihood ratio.[18]

$$\frac{P(E|H)}{P(E|\sim H)}$$

If the likelihood ratio is greater than 1, it is more probable that the agent would have observed the evidence if the hypothesis is true than if it is false. The evidence confirms the hypothesis and the agent should increase her credence in H. If it is less than 1, it is more probable that the agent would have observed the evidence if the hypothesis is false than if it is true. The evidence disconfirms the hypothesis and the agent should decrease her credence in H. If two agents who disagree about H can rationally assign likelihood ratios on opposite sides of 1 to the same set of evidence E, then belief polarization will be a straightforward consequence of Bayesian updating.

This may be the case when agents disagree about auxiliary background beliefs in addition to the target hypothesis, and where the background beliefs affect the expected relationship between the evidence and the hypothesis (and thus the likelihood ratio). For example, suppose that Dr. A has a high credence that a patient has disease D and Dr. B has a low credence that the patient has D. There is a test for D that comes out positive either when the patient has D and also has high blood sugar, or when the patient does not have D and also has low blood sugar. If Dr. A has a high credence that the patient has high blood sugar, while Dr. B has a high credence that the patient has low blood sugar, they will assign a positive test result likelihood ratios on opposite sides of 1 with respect to the hypothesis that the patient has D, and their beliefs about whether the patient has D will polarize (Jern et al., 2014). If we assume that the relevant prior belief distributions are rational, then the Drs. will be rational in polarizing on the test result.

Jern et al. argue not only that rational belief polarization is a theoretical possibility for Bayesian agents in thought experiments like this, but also that many of the empirical results, and, by extension, everyday cases of belief polarization, can be plausibly accounted for in a Bayesian framework. They apply Bayesian models to several empirical studies, among which is the canonical study of polarization of beliefs about the deterrent effects of the death penalty (Lord et al., 1979). In Jern et al.'s model of this study, the agents make two assumptions: That researchers tend to arrive at results that are consistent with their own prior beliefs, and that one's own belief

[18] That is, the probability of observing the evidence conditional on the hypothesis being true, divided by the probability of observing the evidence conditional on the hypothesis being false.

is different from the consensus among researchers. If two such agents initially disagree about whether the death penalty is a deterrent, and also about what the consensus among researchers is, then their beliefs will polarize as a result of Bayesian inference after viewing two conflicting studies. In this case, the background beliefs about researcher bias and the belief that the consensus is wrong warrant taking the study against their prior belief to be a spurious result due to researcher bias, while the study in favor of their prior belief is taken as veridical. Cook & Lewandowsky (2016) use a similar Bayesian model, employing free market support and trust in climate scientists as background variables, to account for their experimental result that beliefs about anthropogenic global warming polarize after subjects view a message conveying that 97% of climate scientists agree that humans are causing global warming. On their model, free market supporters, who are likely to distrust climate scientists, can (in a probabilistically rational manner) regard the consensus message to be evidence against the reality of anthropogenic climate change, reflecting a "fake" consensus manufactured by dishonest scientists in order to deceive.

These Bayesian models are competitors to psychological mechanisms such as selective scrutiny in explaining belief polarization only insofar as they rely purely on prior belief distribution in accounting for biased evaluation of evidence, rather than, for example, the active generation of alternative explanations during selective scrutiny. However, there need not strictly speaking be anything "unbayesian" about such mechanisms. Bayes' rule is, as an updating rule, silent about such matters as whether and how to engage in scrutiny of evidence. Suppose that two hypothetical subjects with opposing prior beliefs in a mixed-evidence experiment do not have prior credence distributions that would allow them to polarize on the basis of the evidence. If however, they engage in selective scrutiny, they thereby generate alternative explanations that alter their credence distributions such that they could now polarize in a manner consistent with probabilistic inference. As long as they ultimately update their beliefs by applying Bayes' rule, there is no genuine conflict between these two accounts.

In spite of the possibility of fitting Bayesian models to the data from belief polarization experiments, there is other empirical evidence that tells against the notion that biased evaluation of evidence is generally the result of Bayesian inference on one's prior belief distribution and the evidence. In studies directly addressing the relationship between selective scrutiny, biased evaluation, and belief polarization, belief polarization is typically found only among those subjects who engage in selective scrutiny and the active generation of counterarguments during evidence evaluation. Thus, it seems unlikely that biased evaluations of the evidence are arrived

at purely through integrating new evidence with background belief in a probabilistic manner (Taber & Lodge, 2006).[19]

So what should we make of the normative implications of the possibility of modeling belief polarization as consistent with Bayesian updating? It is undoubtedly an interesting result that some cases of biased evaluation of evidence and belief polarization can be modeled in this way. However, it ultimately does not constitute a strong reason to think that human agents in typical cases of belief polarization are rational. Assuming that the models accurately capture what is occurring,[20] any normative verdict relies on the further assumption that the relevant prior credence distribution that allows for the biased evaluation of evidence is itself rational. But once we turn away from hypothetical cases such as that of Drs. A and B to more realistic scenarios, it becomes harder to reach this conclusion. Consider again Jern et al.'s (2014) model of the study by Lord et al. (1979). Given their background beliefs, agents in their model were rational in treating the study against their prior belief as a spurious result generated by researcher bias, while treating the study in favor of their belief as veridical evidence. The set of background beliefs that allowed for this was that researchers tend to arrive at results that are consistent with their prior beliefs, and that one's own belief is different from the consensus among researchers. On this background, agents can infer that any study against their belief is likely the result of researcher bias in the direction of the mistaken consensus among researchers. However, any result in favor of one's prior belief must be a veridical piece of evidence, since the researcher bias points in the other direction. But while this set of background beliefs may make such biased evaluation rational in the narrow sense that it does not violate probability theory, it is hard to

---

[19]Jern et al. do mention this study, but only to dismiss its relevance. According to Jern et al., any polarization may have been the result of subjects' self-selecting which evidence to view. However, this is a misunderstanding of the design of Taber & Lodge's study. It did include two tasks: a task of choosing what evidence to view, in addition to an evidence-evaluation task. However, the contents of these two tasks were always related to two different issues, so any polarization that resulted from the evidence evaluation task is unrelated to that resulting from the evidence selection task.

[20] Sometimes this is plainly not the case. Jern et al. (2014), for example, say that no information provided in the studies that they model is inconsistent with their models. However, their model of Lord et al. (1979) is in fact in conflict with information contained in the study. Specifically, Jern et al.'s notion that subjects were assuming that their belief was contrary to the research consensus is flatly denied in the description of how opponents and proponents were identified: "Twenty-four were "proponents" who favored capital punishment, believed it to have a deterrent effect, *and thought most of the relevant research supported their own beliefs.* Twenty-four were "opponents" who opposed capital punishment, doubted its deterrent effect, *and thought that the relevant research supported their views*" (Lord et al., 1979, p. 2100, emphasis mine).

think of a realistic set of circumstances where such a conspiratorial and epistemically arrogant combination of background beliefs would itself be epistemically rational. In discussing their own model, Cook & Lewandowsky (2016) do not consider themselves to have shown that belief polarization about global warming on the basis of a consensus message is rational, as it at least an "open question" whether an expectation of manufactured scientific consensus aimed at deception could itself be considered rational.

3.3 Motivated evaluation of evidence

Kelly's account and the Bayesian models of Jern et al. both construe biased evaluations of evidence as resulting from an accuracy goal: a goal of arriving at an accurate assessment of the evidence and accurate beliefs. For Kelly, individuals engage in scrutiny of evidence against their prior belief because they suspect that, in fact, there are flaws with the evidence, or alternative explanations for it other than the falsity of their belief. It is the suspicion that they would arrive at a wrong evaluation of the evidence if they were to accept it at face value that motivates the scrutiny, and ultimately the biased evaluation of the total set of evidence. Bayesian models see biased evaluations as being the dispassionate assignment of a likelihood ratio to evidence based on one's prior credence distribution. The emphasis on accuracy goals and prior belief as explaining biased evaluation and belief polarization is largely in keeping with early empirical work. Lord et al. (1979), for example, also consider an expectation that evidence for one's belief is likely to be stronger than evidence against it to be the driving factor in biased evaluation and belief polarization.

Psychologists typically contrast accuracy goals with directional goals. When cognition is motivated by a directional goal, it is aimed at reaching a particular, directional conclusion, rather than the most accurate one (whatever that might be). Faced with a cognitive task, the goal of information processing is to yield a construal of the evidence that allows the subject to reach the desired conclusion while maintaining an "illusion of objectivity" – an image of one self as an evidence-driven, objective thinker (Kunda, 1990). Both accuracy goals and directional goals affect how attention is directed, which beliefs are retrieved from memory, and which reasoning strategies are used. But where accuracy goals lead to the use of those mental representations and strategies that are most likely to reach an accurate conclusion, directional goals lead to the use of those mental representations and strategies that are most likely to yield the desired conclusion

(Balcetis & Dunning, 2006; Hennes et al., 2016; Kruglanski & Webster, 1996; Kunda, 1990).[21] To the extent that biased evaluations of evidence occur due to directional goals, the case for their epistemic rationality is considerably weakened. It may be that it can be rational to scrutinize evidence due to an expectation that it must contain flaws. It is another matter to scrutinize it in the hope of finding some pretext to reject it because one would prefer not to change one's mind.

Several studies show that directional goals play a substantial role in explaining biased evaluation of evidence about propositions that are relevant to personal, social, or political values. Kunda (1987) provides an illustrative example. Her subjects, who were either heavy or light consumers of coffee, read an article reporting negative health effects of caffeine consumption. Her expectation was that heavy consumers would be motivated to maintain a self-image as healthy and therefore to rate the study as less convincing than light consumers. However, one might expect that heavy consumers also have different prior beliefs than light consumers about the health effects of caffeine, and the different priors might explain any differences in evaluation. To control for this, the article said that the negative health effects were specific to women. Since there is no reason to suppose that prior beliefs about the health effects of caffeine should vary systematically by gender, any differences in evaluation of the article between male heavy consumers and female heavy consumers could plausibly be attributed to directional goals. The results were in line with that prediction: female heavy consumers rated the article as much less convincing than male heavy consumers, and than low consumers of either gender. Munro & Ditto (1997) conducted two experiments that likewise raise doubts about the role of prior belief and accuracy motivations in explaining biased evaluation. Subjects varied in the degree to which they were generally prejudiced toward homosexuals, and in their degree of belief about a specific negative stereotype about homosexuals. They evaluated two studies, where one confirmed the negative stereotype and the other disconfirmed it. As expected, the study confirming subjects' prejudice (or lack thereof) and prior belief was rated as more convincing, and this resulted in belief polarization. The novelty of their design was that Munro & Ditto also measured subjects' emotional response to receiving each study. Unsurprisingly, studies whose conclusion disconfirmed their prejudice and prior belief tended to evoke a more negative emotional response. More surprisingly, analyses of covariance and a path analysis revealed that, statistically controlling for the effect of emotional response, prior belief did not predict biased

[21] Again, it should be kept in mind that, in spite of talk of goals and strategies, these processes likely occur without conscious awareness.

evaluation at all. General prejudice predicted both prior belief and emotional response, but prior belief itself had no impact on emotional response, or on biased evaluation. In contrast, emotional response predicted biased evaluation. The authors' interpretation is that having their prejudice (or lack thereof) challenged by contrary evidence evoked a negative emotional response in subjects, and selective scrutiny and biased evaluation was initiated in an attempt to avoid this negative emotion. A similar result was found by Munro et al. (2002), whose subjects watched the first 1996 U.S. presidential debate live. Subjects' pre-debate attitude about which candidate they would prefer in office, but not their pre-debate expectation about who would present the best arguments in the debate, predicted biased evaluation of the arguments. Like before, this effect was mediated by emotional response to receiving the arguments.

Studies like this suggest a more general problem: When studies look only at prior belief and subsequent information processing, as is the case for Lord et al. (1979), among others, any association between prior belief and biased evaluation may not be evidence of a causal influence of prior belief on subsequent evidence evaluation, but rather the result of both prior belief and biased evaluation being caused by a third variable. This risk is particularly great for studies that investigate beliefs about controversial topics, as studies on belief polarization tend to do, where third variables such as worldview (Lewandowsky et al., 2013), cultural commitments (Kahan, 2016), or group identity (Dawson et al., 2002) are often highly predictive of biased evaluation. Meanwhile, the true proximal cause of biased evaluation, e.g. emotional responses to evidence, may be missed entirely.

Perhaps the strongest support for such directional goal-accounts comes from studies that demonstrate biased evaluation and belief polarization even when subjects do not differ in their degree of prior belief at all, but do differ with respect to such a background variable. For example, Kahan et al. (2009) measured subjects' cultural commitments[22] and their belief about whether the benefits of nanotechnology exceed its risks. Absent exposure to any novel evidence, there was no difference in prior belief conditional on culture: 61% of both "hierarchical individualists" and "egalitarian communitarians" reported a belief that benefits exceeded risks. However, when subjects were exposed to the same two paragraphs identifying potential risks and benefits of nanotechnology, sharp polarization between the two groups was observed: After observing the evidence, 23% of egalitarian communitarians vs. 86% of hierarchical individualists

[22] Operationalized as location on a two-dimensional scale, where one dimension represents communitarianism vs. individualism, and the other represents support for hierarchical social structures vs. flat, egalitarian ones (Kahan, 2012).

reported a belief that the benefits exceeded the risks. Hierarchical individualists gave much more weight to the evidence of benefits than they did the evidence of risks, whereas the opposite pattern was observed for egalitarian communitarians. Different expectations about the evidence based on different prior beliefs could not be the cause of this differential evaluation, since prior beliefs were identical between the groups.

I have argued that, even operating under the assumption that the picture of biased evaluation sketched by Kelly (2008) and Bayesian modelers such as Jern et al. (2014) is descriptively accurate, their arguments fall short of showing that biased evaluation in cases where it causes belief polarization is rational. We now see that the descriptive picture is itself unlikely to be accurate in such cases, and should be replaced by one on which it is much harder to construct plausible accounts of why such biased evaluation of evidence is rational (Strickland, Taber, & Lodge, 2011).[23]

4 Is polarized belief responsive to evidence?

We now turn to the second normative question: Can the polarized beliefs that result from biased evaluations of evidence be rational? At a first glance, the answer might seem obvious if we assume that the negative verdict we arrived at with respect to biased evaluation is correct. If beliefs are based on irrational reasoning, are those beliefs themselves not obviously irrational? But it is worth considering the question more carefully. Even if the reasoning itself is irrational, in the sense that it is hard to construct a rational account of the cognitive processes that actually shape evaluations of evidence, subjects might not be aware that they are doing anything odious. Might they then not be rational in updating their belief on their interpretation of the evidence, even if their evaluation of it is biased? Recall also that on Kelly's view, the notion of rationality at work in evaluating biased evaluation is, in a certain sense, practical. Scrutinizing evidence is an action, competing for time and resources with other potential actions. In the final analysis, whether one is rational in scrutinizing a piece of evidence may depend on matters such as whether one is in a hurry to make an appointment. In contrast, the epistemic rationality of beliefs

[23] At least for Kelly's picture and Bayesian models that only rely on prior credence distributions. Some Bayesian models, including the one used by Cook & Lewandowsky (2016), include variables such as worldview. But, as Cook & Lewandowsky seem to agree, the ability to model how worldview impacts evaluation of evidence does not make that impact more rational.

is typically not taken to be sensitive to such matters, but solely as a question of whether they are supported by one's total evidence, or are the output of a reliable belief-forming mechanism. Given these different standards of evaluation, it is a possibility that the verdicts need not be in the same direction for biased evaluation and its resulting beliefs.

While Lord et al. (1979) do not unequivocally condemn their subjects' biased evaluation of the evidence, they are more clear with respect to the resulting polarized belief: "[The subjects'] sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially "justified" the processing bias. In so doing, subjects exposed themselves to the familiar risk of making their hypotheses unfalsifiable—a serious risk in a domain where it is clear that at least one party in a dispute holds a false hypothesis." (p. 2107). For philosophers with reliabilist inclinations, it might also seem an easy conclusion that polarized beliefs based on a biased evaluation of evidence are unjustified. Necessarily, in a case of belief polarization, one person ends up more certain of a falsehood. Biased evaluations of evidence will tend to pull one further towards certainty of a falsehood whenever one's desired conclusion is false. Of course, the converse is also true: If one is motivated to reach a conclusion that is in fact true, then motivated reasoning will increase the likelihood that one comes to believe a truth, compared to even-handed processing (Dawson et al., 2002; Kahan, Peters, et al., 2017). Nevertheless, where an unbiased evaluation of evidence will always be truth-conducive insofar as one is competent and the evidence itself is truth-tracking, biased evaluation will be truth-tracking only insofar as one's desired conclusions happen to be true. For a reliabilist, this might seem sufficient for biased evaluation to fail to be a reliable process, and so for any polarization of beliefs produced by it to be unjustified.

However, Kelly (2008) offers a subtle account of why polarized beliefs based on biased evaluations of evidence can be rational.[24] Selective scrutiny results in the discovery of alternative explanations of evidence for which the falsity of one's belief is a potential explanation. Awareness of these alternative explanations does not only play a causal role in subjects' decreasing the weight that they give to evidence against their prior beliefs. From a normative

[24] It is important to note that Kelly's defense assumes that subjects are unaware that their evaluation of evidence is biased. Much of the latter portion of his discussion focuses on what should happen once subjects learn about biased evaluation, and he argues that such awareness would give one reason to correct for the role that bias has played in skewing the sample of considerations one has available. My disagreement here is with Kelly's verdict about rationality when subjects are unaware of the impact of bias.

standpoint, the alternative explanations should themselves be counted among the subjects' evidence, broadly construed. Polarized belief can be a rational reflection of one's total evidence, including any such alternative explanations.[25] This brief sketch of his argument deserves some elaboration. What does the work is what Kelly (2008, p. 620) refers to as the Key Epistemological Fact:

"For a given body of evidence and a given hypothesis that purports to explain that evidence, how confident one should be that the hypothesis is true on the basis of the evidence depends on the space of alternative hypotheses of which one is aware."

As an example of the Key Epistemological Fact at work, Kelly notes that the credence one could rationally afford the Design Hypothesis of biological complexity drastically diminished with the advent of the theory of evolution. Prior to Darwin, the Design Hypothesis was, plausibly, the sole explanation available for biological complexity, and therefore a high credence in Design was warranted based on the available evidence. With the theory of evolution by natural selection, people were suddenly aware of an alternative explanation for biological complexity. Awareness of this alternative explanation meant that, even disregarding the direct evidence in favor of evolution, the rational credence that one could lend to Design based on the evidence of biological complexity diminished.

Now consider how this applies in a case of belief polarization. Suppose that one reads a study whose results are seemingly at odds with one's prior belief. One hypothesis explaining the result is that one's prior belief is false. But upon engaging in scrutiny of the study, one might uncover what one takes to be flaws in the study. If so, one becomes aware of alternative hypotheses that might explain the result, aside from the falsity of one's belief. The Key Epistemological Fact tells us that these alternative hypotheses make it rational for us to take the study to count against our belief to a lesser degree than it would have in their absence, since the falsity of our belief is only one of several hypotheses that might explain the results of the study.

---

[25] Kelly emphasizes these points in relation to a discussion of whether the impact of prior belief on subsequently evaluated evidence means that resulting beliefs violate a principle of the commutativity of evidence. The principle says that, for a total body of evidence, the temporal order with which one receives its parts cannot make a difference for what one is ultimately rational in believing. Kelly argues that biased evaluation of evidence means that the temporal order with which one receives evidence does make a *causal* difference to what alternative explanations one has in mind, and since these alternative explanations are part of one's total evidence, it is not the case that polarized belief constitute violations of the commutativity principle.

On the other hand, when assessing a study whose results count in favor of a prior belief, we are unlikely to have such alternative explanations available given the relative lack of scrutiny, and the tendency for any engagement with the study to provide bolstering, rather than denigrating, reasons (Taber & Lodge, 2006). As such, we are not aware of any reason to take the supportive study on board in anything less than full force. Naturally, Kelly does not claim that any alternative explanation can justify decreasing the probative weight of contrary evidence. To the extent that alternative hypotheses are ad hoc or implausible, a belief based on disregarding a study on their basis would be irrational. But if one succeeds in discovering plausible alternative explanations, such as genuine problems with methodology, analysis, or argument, then polarized belief is a rational response to one's total evidence.

4.1 Normative defeat

There is much here to agree with. Something like the Key Epistemological Fact seems plausible. If one has found what is actually a flaw in study that harms its validity, then it seems clear that one is rational in decreasing the weight given to that study in one's posterior belief. Still, Kelly's account does not seem entirely satisfactory. In particular, his verdict with regard to evidence in favor of our prior beliefs strikes me as lacking. Kelly holds that, if one is unaware of flaws in supporting studies, and therefore has no alternative explanations in mind other than the truth of one's belief, one is justified in giving full weight to such studies in bolstering one's belief. But this seems problematic. Selective scrutiny can just as easily be described as selective laziness (Trouche et al., 2016). When faced with evidence for which the truth of one's belief seems to be a potential explanation, individuals are unlikely to engage in the kind of critical scrutiny that might uncover plausible alternative explanations of the data, or even reveal them to be flat-out mistaken about the bearing of the evidence on the belief (Dawson et al., 2002; Kahan, Peters, et al., 2017).

In order for both subjects to be rational in a standard case of belief polarization on Kelly's view, both parties must find plausible alternative explanations for the evidence against their prior belief.[26] If such plausible alternative explanations were only available to be discovered by one side, the other could not reasonably discount the counterevidence and use the total evidence to bolster his or her prior belief. The upshot is that what makes polarized belief rational in such

[26] This is at least the case in the mixed-evidence cases like Lord et al. (1979) where the data, taken in isolation, is equally probative on both sides.

cases is as much a failure to discover what are in fact plausible alternative explanations for the evidence that supports one's view as it is the successful discovery of alternative explanations for the data against one's view. Furthermore, it is not the case that discovering the alternative explanations for the supporting evidence is beyond one's abilities. If it were, then one would also not have been able to discover the flaws in the opposing studies of a similar design and quality. The failure is due to selective laziness. Therefore, we can say that polarized beliefs in such cases are generally beliefs that would not hold up to one's own critical scrutiny. And it at least sometimes taken as a requirement of rational belief that it would hold up to such scrutiny (BonJour & Sosa, 2003).

We can construe the disagreement I have with Kelly more carefully in the language of defeaters. On Kelly's view, what justifies polarized belief is that one downgrades the probative force of studies against a prior belief in the light of doxastic defeaters: Beliefs held by the agent, whose truth would make it improper to fully accept the conclusion of the study. Since there are no doxastic defeaters in the case of supporting evidence, there is no reason not to fully accept the conclusion of the study.

But doxastic defeaters are not the only type of defeater that has been discussed by philosophers. Another type of defeater worth mentioning in the present context is the normative defeater. A normative defeater is a proposition q that a subject ought to believe (whether or not they actually believe it), which indicates that the subject's belief that p is false (Lackey, 2005). Normative defeat implies that propositions one is unaware of can be part of one's total evidence, in the broad sense that they make a difference to how rational one's belief is. But this is something that Kelly (2008, p. 630) explicitly denies: "In general, accurately proportioning one's beliefs to one's total evidence suffices for believing reasonably. But facts of which one is completely unaware are not eligible for inclusion among one's total evidence."

This construal of evidentialism is at least moderately controversial. Compare, for example, to Kopec's construal (2017, p. 14): "[Evidentialism], in its pure form, holds that an agent ought to believe a proposition if and only if that proposition is supported by the agent's total evidence, and this biconditional is supposed to hold even if the agent is not aware of the evidential relations at issue (Conee & Feldman, 2004). On some versions of the view, it is supposed to hold even if she justifiably believes false things about the evidential relations." The fact that plausible alternative explanations exist means that the actual evidential relation between, say, a study and a hypothesis, is weakened. Individuals who fail to discover these alternative explanations are polarizing their belief on the basis of a mistaken view of the evidential relation, rather than the

actual evidential relation. So on this version of evidentialism, agents ought not polarize in typical cases, because in so doing they are not basing their views on the actual evidential relations at issue, but on their misperceptions of the evidential relations.

There are cases where the rationality of a belief clearly appears to depend on evidence of which the subject is unaware. Suppose that your doctor believes, on the basis of his or her medical training, that treatment X is best practice for a certain common disease that you have contracted. In fact, over the last few years, ample evidence has accumulated in the medical literature showing that treatment X is inferior to treatment Y. Your doctor is not aware of this evidence, since he or she has not bothered to keep up with the medical literature. It seems natural to say that the evidence in the medical literature constitutes a normative defeater of the rationality of your doctor's belief that treatment X is best practice. His or her belief is less than perfectly rational, even if he or she has responded perfectly to the total evidence of which he or she is aware.[27]

It is likewise plausible that there are normative defeaters at play in typical cases of belief polarization. Specifically, the undiscovered alternative explanations for supporting evidence may be propositions that subjects ought to believe, and that count against the truth of their belief that the truth of their prior belief is the actual explanation of the evidence. Suppose that one evaluates a study, finds no plausible alternative explanations, and concludes that the truth of one's belief is the actual explanation of the results of the study. To the extent that there are in fact alternative explanations that one would have found if only the results had pointed in the other direction, and so one would not have been lazy, polarized belief that results from missing the alternative explanations are subject to normative defeat. Believers ought to be aware of such alternative explanations, even if they are not actually aware of them.

It is important for the plausibility of this conclusion that the alternative explanations are in fact available to subjects, in the sense that only selective laziness prevents their discovery. Compare to the case of the Design Hypothesis. It is presumably not the case that most people prior to Darwin had the competence to think up evolution by natural selection as a plausible alternative explanation of biological complexity, but merely lacked the motivation. So here there does not seem to be a normative defeater for taking Design as the explanation of biological complexity. The fact that subjects in the belief polarization experiments easily *could* have

[27] This case is inspired by cases presented by Goldberg (2017), who argues that there are cases where subjects should have known that p in spite of lacking evidence for p.

discovered the flaws if only their motivation inclined them to do so makes it more plausible that there are normative defeaters in these cases.

To press this point a bit further, consider cases of belief polarization where selective laziness leads not to a failure to discover alternative explanations for somewhat ambiguous pieces of evidence, but to misconstruing evidence that actually, and unambiguously, disconfirms one's prior belief as being confirmatory. In a study by Kahan et al. (2017), subjects were presented with the results of a study comparing crime trends between two groups of cities. One group had recently enacted a ban on carrying concealed handguns in public, and one did not have such bans. Subjects were shown the results in the following form. Numbers in each cell indicate the number of cities observed.

	Crime increased	Crime decreased
Cities that did ban carrying concealed handguns in public	223	75
Cities that did not ban carrying concealed handguns in public	107	21

So, out of the total of 426 cities in the study, 298 had enacted a ban, and 128 had not. Out of the 298 cities that had enacted the ban, crime increased in 223 and decreased in 75. Out of the 128 cities without the ban, crime increased in 107 cities and decreased in 21.

Subjects were then asked which of the following conclusions the study supported:

1. Cities that enacted a ban on carrying concealed handguns were more likely to have a decrease in crime than cities without bans.
2. Cities that enacted a ban on carrying concealed handguns were more likely to have an increase in crime than cities without bans.

The correct answer in the above version is that crime was more likely to decrease in cities that implemented the ban compared to cities that did not (for half the subjects, the "crime increased"

and "crime decreased" cells were swapped, such that the opposite conclusion was correct). In tasks like this, most subjects have an immediate intuition about what the evidence supports, but this intuition pulls in the wrong direction (Wasserman, Dorner, & Kao, 1990). Finding the correct solution requires first overriding this intuition, and then engaging in the deliberate reasoning required to find the correct response (Evans & Stanovich, 2013). Subjects whose political ideology was threatened by the intuitive, but false, interpretation of the data were motivated to engage in reasoning to reject the intuitive response, and were much more likely to answer correctly than controls.[28] However, subjects whose political stance was affirmed by the intuitive, but wrong, interpretation of the data were motivated to be "lazy" and accept the intuition at face value. They were therefore much more likely to answer incorrectly than controls.

So these subjects concluded that the evidence supported the *opposite* conclusion of what it did in fact unambiguously support. In a certain sense, bolstering their prior belief on this basis would simply be conforming their belief to the evidence they are aware of: an intuitive sense that the data supports their politically favored conclusion. But it seems clear that this would be irrational. The fact that the evidence unambiguously supports the opposite conclusion, combined with the fact that subjects would likely have discovered this if only their intuition supported the opposite conclusion, defeats the rationality of their bolstered belief, even though subjects are likely unaware that they have misinterpreted the data.

It is not clear that there is a reason to treat differently unambiguous evidence and the more ambiguous mixed evidence in typical belief polarization studies. If the probative force of a study is in fact severely diminished by plausible alternative explanations that should easily be discovered, then unawareness of the alternative explanations due to selective laziness does not make a bolstered belief based on giving that study full weight rational.

This discussion leaves us with the following modified version of the Key Epistemological Fact: "For a given body of evidence and a given hypothesis that purports to explain that evidence, how confident one should be that the hypothesis is true on the basis of the evidence depends on the space of alternative hypotheses of which one is, *or ought to be*, aware." In standard cases of

[28] Liberals were threatened by the conclusion that gun control increases crime, while conservatives threatened by the conclusion that gun control decreases crime. Controls solved an identical but politically neutral task about the effects of a novel skin crème for treating a rash.

belief polarization, the polarized beliefs are irrational because selective laziness causes one to disregard alternative hypothesis that one ought to be aware of.

5 Concluding remarks

In society at large, beliefs about many matters of fact that are of great importance to human well-being are polarized along political, cultural, or social fault lines. One prominent reason for this is that individuals are likely construe evidence as supportive of their desired conclusions. I have argued that biased evaluations of evidence in such contested domains, and the resulting polarized beliefs, are not rational. This is the case whether evaluations of evidence are biased by expectations based on prior belief or, as is more often likely the case, by desires to reach congenial conclusions.

The upshot is that many of our beliefs about publically contested issues are likely to be less rational than we would like. We should be worried about belief polarization, both our own and in society at large, not just for pragmatic or political reasons, but also for purely epistemic reasons.

Article 2: The epistemic significance of political disagreement

1 Introduction

The epistemic significance of disagreement has become a widely debated issue in epistemology. While there is general agreement that the kinds of disagreement we encounter in real life can be of epistemic significance,[29] most of this debate has focused on the idealized case of peer disagreement. There are two widely employed notions of epistemic peerhood in the literature.

Kelly (2005, p. 174), states that two individuals are epistemic peers with respect to a question if and only if they satisfy the following two conditions:

(i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question, and
(ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness, and freedom from bias.

What does 'equal' mean here? Starting with (i), some philosophers suggest that to be equals with respect to evidence is to have identical evidence (King, 2012). However, I think there is good reason to think that what is epistemically significant, in the sense of determining the amount of doxastic change required, is not equality in the sense of identity, but equality in the sense of parity. Suppose that you and I discover that we disagree about p, and that I have good reason to think that while your evidence about p differs from mine, it is just as good. This seems to provide me with as much reason to doubt the truth of my belief as your disagreeing on the basis of identical evidence does (Christensen, 2007).[30] Similar considerations apply to (ii). What is

[29] See e.g. Matheson (2015), Christensen (2014b), Lackey (2008b), and King (2012).

[30] One oft-cited reason why peer disagreements are epistemically significant is that they raise the possibility that one of the peers has made some mistake in their evaluation of the evidence. This possibility is certainly more salient in cases of identical evidence. But Christensen's case shows that it would be mistaken to think that this is the only kind of worry that peer disagreements can raise. To the extent that one's interest is investigating the epistemic significance of the kind of higher-order evidence of a cognitive malfunction that arises to a greater extent when there

epistemically significant is that the two parties to a disagreement are equally virtuous, in the sense that they are generally equally good at responding to the evidence – not that they respond to evidence in identical ways (Matheson, 2015). So, roughly, two people are epistemic peers in the first sense if there is evidential and cognitive parity between them.

The second widely used notion of peerhood is from Elga (2007). On Elga's view, you count someone as your epistemic peer "…with respect to an about-to-be-judged claim if and only if you think that, conditional the two of you disagreeing about the claim, the two of you are equally likely to be mistaken" (2007, p. 499). So, I consider you to be my epistemic peer with respect to p if I think that, should it turn out that we disagree about p, it is as likely that your belief is mistaken as it is that mine is. As I understand Elga, being mistaken does not simply mean that your belief is false, nor does being right simply mean that your belief is true. Rather, the notions of mistaken and right reflect whether one has adopted the doxastic attitude that is rational in light of the available evidence. Suppose that I consider you an absolute expert with respect to the proposition that it will rain tomorrow, in the sense that conditional on our disagreeing, I give it probability 1 that you are correct. Your credence in the proposition that it will rain tomorrow is .6, so I also adopt credence .6 when I learn of your opinion. Of course, either it rains tomorrow or it does not. But I take your doxastic attitude to be 'right' in the sense that a credence of .6 is the right doxastic attitude to adopt with respect to the proposition that it will rain tomorrow, in light of the available evidence on that matter.

These two notions of epistemic peerhood are, clearly, closely related. When I think that you are in possession of equally good evidence and that you are equally good at evaluating the evidence, then I am very likely to also think that, should it turn out that we disagree; the probability that I am right is the same as the probability that you are right. Indeed, Elga takes his notion of peerhood to correspond to the proposition that your interlocutor is "as good as you at evaluating claims on the matter" (Elga, 2007, p. 484). But two people being equally good at evaluating claims looks quite similar to there being cognitive parity and evidential parity between them.

---

is evidential identity, one can add to epistemic peerhood a stipulation that there has been full disclosure of the evidence (Christensen, 2010; Feldman, 2006; Lackey, 2008b).

There is, appropriately enough for the topic, much disagreement among philosophers about what the best notion of epistemic peerhood is (Gelfert, 2011), and about what doxastic revision, if any, is required when an individual discovers that they are party to a peer disagreement (Christensen & Lackey, 2013; Feldman & Warfield, 2010). What there does not seem to be any particular disagreement about is that the factors captured in the two notions of peerhood are important determinants of how much doxastic revision is required in a case of disagreement. When I find myself in a disagreement with someone whom I think is clearly my cognitive and evidential superior, there is widespread consensus that I should defer to his or her judgment. I should do the same if I find myself in a disagreement with someone whom I, antecedently to discovering the disagreement, thought much more likely to be correct about the matter than myself. At the other end of the spectrum, it is not very controversial that I should not be moved when I disagree with someone who has absolutely no evidence about the disputed matter, but I have great familiarity with much high-quality evidence. Neither should I be moved when I disagree with someone who is much, much worse than me at evaluating any evidence he or she might possess. Likewise, if I think that the probability that you are right conditional on our disagreeing is 0, I should not be moved by the discovery that we disagree.

What this shows is that the extent of doxastic revision one's evidence requires in a case of disagreement is, at least in part,[31] a function one's interlocutor's evidence, cognitive faculties, and accuracy. The intuition that accuracy strongly correlates with the combination of evidential quality and cognitive quality is, I think, the reason why most discussions in the epistemology of disagreement focus only on one of these factors. Nevertheless, we will need to keep track of both notions for the remainder of this paper. For ease of reference, we can name the two notions as follows:

ABILITY: An individual's familiarity with the relevant evidence and arguments, and their epistemic virtues such as intelligence, thoughtfulness, and freedom from bias.

ACCURACY: An individual's probability of being correct, conditional on their disagreement with another individual.

[31] I do not wish to exclude the possibility that possibility that factors such as who actually got things right, or whether a belief is actually justified or not can also make a difference (Kelly, 2010, 2013; Lackey, 2008b).

So in cases of disagreement, holding any facts about myself constant, the higher I rate my interlocutor's ABILITY or ACCURACY, the more weight I should give to their disagreement.

In what follows, I show that one's beliefs about an interlocutor's ABILITY and beliefs about their ACCURACY should be inversely correlated in cases of disagreement about politically charged facts. In such disagreements, the more familiar I believe you are with the relevant evidence, and the more reflective, intelligent, numerate, open-minded, scientifically literate, and well-educated I think you are, the more likely I should think it is that you are wrong. This presents us with a conundrum when trying to estimate the epistemic significance of our disagreement.

2 Motivated reasoning and political polarization of factual beliefs

Our doxastic attitudes toward politically charged propositions, such as *humans are causing global warming*, or *the death penalty is a deterrent to murder* are likely to be influenced by motivated reasoning (Corner, Whitmarsh, & Xenias, 2012; Kahan et al., 2012; Lord et al., 1979). By motivated reasoning, I mean reasoning that is aimed at arriving at an interpretation of evidence that yields a desired conclusion – in this case, the conclusion that vindicates one's political views (Kahan, 2016; Kunda, 1990; Van Bavel & Pereira, 2018). When under the influence of motivated reasoning, we (unbeknownst to ourselves) utilize our cognitive abilities to generate reasons in favor of a desired conclusion, and to fend off reasons to against it. The greater one's ABILITY, the better one is at making the evidence yield the desired conclusion (Kahan, 2013; Kahan, Peters, et al., 2017; Kraft et al., 2015).

Motivated reasoning explains an otherwise strange fact about the distribution of doxastic attitudes about politically disputed propositions in the general population: Not only are our beliefs polarized along political fault lines, such that political ideology is highly predictive of one's belief about these propositions, but the degree of polarization correlates positively with familiarity with relevant evidence, and with measures of cognitive ability, numerical reasoning skills, education, scientific literacy, and even with intellectual virtues such as open-mindedness and reflectiveness (Hamilton, 2011; Kahan et al., 2012; Kahan & Corbin, 2016; Kahan & Stanovich, 2016; Taber & Lodge, 2006). In other words, the greater a person's ABILITY, the more predictive that person's political ideology is of their beliefs about politically charged propositions, and the more likely they are to adopt extreme doxastic attitudes. To give an example, liberal democrats are likely to believe that humans are causing global warming, and

their likelihood of believing this, and the certainty with which they hold the belief, increases with their familiarity with relevant evidence, education level, ability to inhibit intuitions and reflect on complex problems, their open-mindedness, scientific literacy, etc. In contrast, conservative republicans are likely to disbelieve that humans are causing global warming, and their likelihood of disbelieving it and their certainty in their disbelief increases with familiarity with relevant evidence, education level, ability to reflect, open-mindedness, scientific literacy, etc. (Hamilton, 2011; Kahan et al., 2012; Kahan & Corbin, 2016).[32]

3 The conflict between ABILITY and ACCURACY

The above findings suggest that there is a conflict between ABILITY and ACCURACY in determining the epistemic significance of disagreements about politically divisive propositions. Consider the following case:

BRILLIANT PARTISAN: I belong firmly on one side of the political aisle. I have a doxastic attitude toward a politically disputed proposition p that is typical for those on my side of the aisle. I learn that you are an extremely intelligent, highly educated and scientifically literate, open-minded, and reflective person on the other side of the political aisle, who is intimately familiar with the relevant evidence about p.

How epistemically significant should I consider any subsequent evidence that you disagree with me about p? If I assess the epistemic significance of our disagreement by reference to your ABILITY, then it would seem that I should find it quite significant. After all, you score highly on virtually all the defining qualities of ABILITY. But if I assess the epistemic significance of our disagreement with respect to ACCURACY, then it seems that I should find our disagreement relatively less significant. Given what I know about you, I think it highly likely that you have adopted an extreme doxastic attitude toward p in the opposite direction of my own. I think that it is very unlikely that your belief would be correct if this were the case – I just couldn't fathom

[32] This pattern is even stronger when, rather than looking at political partisanship or ideology, one looks at individuals' cultural commitments – the degree to which they prefer market-driven, bottom up solutions to social problems over collective and top-down ones, and the degree to which they prefer relatively stratified social organizations with clear power- and status differentials over more egalitarian ones (Kahan, 2012).

how the publically available evidence about p could support a doxastic attitude like that. So, conditional on our disagreeing, I think it very unlikely that you are correct.

So, ABILITY and ACCURACY yield different verdicts about the epistemic significance of this case. But it gets worse. Consider the following variation on the case:

MEDIOCRE PARTISAN: I belong firmly on one side of the political aisle. I have a doxastic attitude about a politically disputed proposition p that is typical for those on my side of the aisle. I learn that you are a moderately intelligent, decently educated and scientifically literate, somewhat open-minded, and reasonably reflective person from the other side of the political aisle, who has some familiarity with the relevant evidence.

With respect to ABILITY, I should give less weight to the disagreement in MEDIOCRE PARTISAN than I did in BRILLIANT PARTISAN. This much seems clear: There is no intellectual virtue that you have to a greater extent than you did in BRILLIANT PARTISAN, and there are several that you have to a lesser extent, including lesser familiarity with the evidence. But on ACCURACY, it can be argued that I should give more weight to the disagreement in MEDIOCRE PARTISAN than I did in BRILLIANT PARTISAN. In light of your lesser virtuousness, you are more likely to have taken a more moderate doxastic attitude toward p. I might even think it somewhat likely that your doxastic attitude toward p leans in the same direction as my own, since your political position is less predictive of your substantive view. Either way, I find it much more plausible that the publically available evidence about p could support a moderate doxastic attitude in the opposite direction of my own than I find it that it could support an extreme attitude in that direction. So, I might think that it is more likely that you are right in MODERATE PARTISAN than I do in BRILLIANT PARTISAN. If so, then the two factors that are often thought to be coextensive, or at least strongly positively correlated, are in fact inversely correlated in political disagreements.

So, what factor should we base our assessment of the epistemic significance of the disagreement on: ABILITY or ACCURACY? Neither verdict seems entirely satisfactory. If we go with ABILITY, then I should give more weight to the disagreement the more likely I think you are to be wrong. If we go with ACCURACY, I should give more weight to the disagreement the less familiar I think you are with the relevant evidence, and the less competent I think you are in processing it.

But maybe there's some mistake here. After all, one of the defining qualities of ABILITY is freedom from bias. As you are subject to motivated reasoning, you are not free from bias. And the evidence on motivated reasoning might appear to suggest that you're *more* biased in BRILLIANT PARTISAN than you are in MEDIOCRE PARTISAN. Perhaps this higher amount of bias cancels out, or even outweighs, the higher levels you possess of the other qualities. So taking the total qualities into account you do not, in fact, have higher ABILITY in BRILLIANT PARTISAN than you do in MEDIOCRE PARTISAN.

However, there are problems with this solution. As it relates to motivated reasoning, freedom from bias means that one's reasoning does not proceed with the goal of defending a particular conclusion, but rather proceeds with the goal of arriving at an accurate assessment of the evidence (Kunda, 1990). But in this sense, you are equally biased in BRILLIANT PARTISAN and MEDIOCRE PARTISAN. In both cases, you are equally motivated to arrive at the conclusion that is congenial to your political outlook. It is just that you are more successful at doing so in BRILLIANT PARTISAN, because your other qualities are put to use in construing the evidence so as to support the desired conclusion, and you have an advantage with respect to those other qualities. Of course, we might take freedom from bias to indicate something else. We might take it to indicate that your assessment of the evidence yields a construal that is reflective of what the evidence actually supports. But doing so simply equates freedom from bias with ACCURACY. And this doesn't seem to be what is normally meant by freedom from bias. What is meant is that one's reasoning is free from any systematically biasing factors, not that we necessarily get things right. The biasing factor in question here is a motivation to arrive at a desired conclusion, and you have that motivation to equal extent in the two cases.

Another mistake that might snuck its way in concerns how we construed the verdict yielded by assigning epistemic significance based on ACCURACY. Am I really justified in thinking it less likely that you are correct in BRILLIANT PARTISAN than in MEDIOCRE PARTISAN, merely on the basis that your view is farther removed from my own? The reasoning driving that verdict might be something like the following: Suppose that, prior to learning anything about you, I were to reflect on the possibility that the evidence about p really supports some other credence than what I currently assign p. As a result, I form a credence distribution over what credence the evidence really supports about p. Plausibly, this credence distribution should have a maximum at the credence I currently assign p. If I didn't think it more likely that the evidence supports my current credence in p than that it supports some other credence, then I wouldn't have my current credence in p at all. But what probability should I assign other credences as we move away from

my actual one? Plausibly, I shouldn't assign all other credences equal probability. It would be odd for me to think that it is just as likely that the evidence really supports a credence that is as far from my current credence as possible as it is that it supports a credence in the very near vicinity of my current one. So let's say that my credence distribution should be something like a normal distribution around my actual credence in p. The farther away from my actual credence we move, the less likely I think it is that the evidence really supports that credence. If so, it seems like I have a reason to think that it is less likely that someone with a credence in p that is very far from my own is correct, compared to someone with a credence closer to my own.

But this line of reasoning runs afoul of the Independence principle (Christensen, 2009). This principle says that I should not use my initial reasoning about p to assess your epistemic credentials when determining the epistemic significance of disagreement. I can only assess it by referring to dispute-independent reasons. But my higher-order credence distribution is based on my initial reasoning about p: it has the shape it does because I initially judge that p is the right response to the evidence. So I cannot use it to assess your ACCURACY.

However, there is another route to the conclusion that you have less ACCURACY in BRILLIANT PARTISAN than you do in MEDIOCRE PARTISAN. Your brilliance means that you are extremely good at coming up with reasons in favor of your desired conclusion and at generating justifications for rejecting any reasons you should encounter against it. But your ability to come up with all these reasons is not very informative about what the evidence actually supports with respect to p. Your brilliance means that, more or less regardless of what the evidence really supports about p, you would be able to conjure up reasons to be supremely confident that your desired conclusion about p is correct. Even with respect to a piece of evidence that strongly suggests that you are wrong about p, you would be able to creatively wriggle your way out of taking that evidence as counting against p. Not so for you in MEDIOCRE PARTISAN. When faced with strong evidence against your favored conclusion here, you do not have the ABILITY necessary to escape from the conclusion that the evidence really warrants decreasing your confidence in your favored conclusion about p. So your belief about p is more constrained by what the evidence actually supports in MEDIOCRE PARTISAN than it is in BRILLIANT PARTISAN. Therefore, I should think that you have higher ACCURACY in MEDIOCRE PARTISAN than you do in BRILLIANT PARTISAN. This line of thought makes no reference to my own reasoning about p. It only makes use of the general claim that, due to your brilliance, you are generally better able to shape your interpretations of the evidence in ways that are congenial to your favored conclusions.

4 Introspection and symmetry

If the above line of reasoning is correct, then we should think that there is an inverse correlation between ABILITY and ACCURACY in cases of disagreement about politically disputed propositions. Choosing to base our epistemic assessment on either of these leads to an uncomfortable conclusion. But perhaps it doesn't matter whether we go with ABILITY or ACCURACY in determining the epistemic significance of such disagreements, because we should not consider such disagreements to be epistemically significant at all, regardless of what standard we go with.

We can use a type of debunking argument to get that conclusion. Your being biased in both BRILLIANT PARTISAN and MEDIOCRE PARTISAN gives me a reason to disregard your disagreement with me altogether. Why should I take your belief about p into account at all if I have good reason to suspect that your have arrived at the belief through biased reasoning about the evidence? In the domain of philosophical disagreement, Fumerton (2010) has argued that he often takes himself to have good reason to discount the disagreement of his colleagues, even well-known and respected ones, because he takes them to be subject to motivated reasoning: "Perhaps I have some reason to believe, for example, that [my disagreeing colleagues] are the victims of various biases that cause them to believe what they want to believe. Indeed, I suspect that I do have reason to believe that others are afflicted in such ways…" (2010, p. 102). So maybe it is not hard at all to determine the epistemic significance of disagreement about politically charged propositions. We should not take them to be significant at all, because we often have dispute-independent reason to think that the views of our interlocutors are the result of motivated reasoning.

But this proceeds too quickly. After all, are our own beliefs not similarly afflicted? What reason do I have to suppose that my own belief about p is not the result of my political ideology having influenced the way I have sought out and evaluated the relevant evidence? Is there a reason to think that those on the other side of the political aisle are generally subject to motivated reasoning to a greater extent than those on my side of the aisle? Not according to the best relevant research. Liberals and conservatives are equally motivated to arrive at politically congenial conclusions, and are equally likely to fall prey to similar magnitudes of motivated reasoning (Frimer et al., 2017; Hallen et al., 2017; Kahan, 2013). Therefore, one's political position can provide no independent reason to take oneself to be absolved from the charge of bias, at least not on our current evidence about the political distribution of motivated reasoning.

While the studies showing political symmetry in the degree of motivated reasoning might apply

in general, they only show that on average, liberals and conservatives are equally biased. This of course leaves open the possibility that I am not biased. Could I not therefore have some reason to think that I, unlike most other people on both sides of the political aisle, am exempt from bias? While I think that statistical information of this nature should generally influence our beliefs about ourselves (Elga, 2005), let us nevertheless explore that question. Fumerton appeals to self-knowledge gained through introspection to argue that he has reason to think that he is not subject to bias to the same extent as other philosophers: "I do, in fact, think that I have got more self-knowledge than a great many other academics I know, and I think that self-knowledge gives me a better and more neutral perspective on a host of philosophical and political issues" (Fumerton, 2010, p. 102). Following this line of thought, Fumerton takes himself to be justified in downgrading the epistemic significance of philosophical disagreement by attributing motivated reasoning to his colleagues and exempting himself through introspective evidence of the absence of bias. So perhaps we can do the same in cases of political disagreement.

No, we cannot, for reasons similar to those presented by Ballantyne (2015). Among the many documented biases in human cognition is the so-called bias blind-spot (Pronin et al., 2002). The bias-blind spot refers to a sort of meta-bias, a broad tendency to readily attribute biases to others, while considering one's own reasoning to be objective and neutral, even when one is in fact biased. Bias blind-spot, and other research indicating that introspection is generally an unreliable process of generating beliefs about one's own cognitive processes (Carruthers, 2011; Nisbett & Wilson, 1977), suggests that Fumerton is not justified in attributing motivated reasoning to other philosophers but not himself on the basis of introspection. So it goes for political disagreement. Bias blind-spot implies that one is not justified in using attributions of motivated reasoning as a symmetry breaker unless one has independent reason outside of introspection to exculpate oneself. It might be true that you are subject to motivated reasoning, but I have no independent reason to suspect that I am not subject to the same, even if introspection seems to suggest that I am not.

Someone arguing along Fumerton's lines might reply that his substantial self-knowledge includes his knowledge that bias blind-spot and motivated reasoning is a risk in his own thinking about politically disputed propositions. In cases of disagreement, this awareness causes him to take extra care to make sure that his reasoning is unbiased. Since he knows that he has sincerely attempted to correct for bias, but he does not know that you have done so, he has personal information that can serve as a relevant symmetry breaker (Lackey, 2008b). But this reply is unsatisfactory. In fact, a general knowledge of the existence of motivated reasoning and bias blind spot can put one at risk of succumbing to the bias to an even greater extent than one would

in the absence of such knowledge. Suppose that you reason from some evidence to p. Because of your awareness of the existence of motivated reasoning, and of the blind spot that makes its operation opaque to yourself, you make sure to double-check that your reasoning is strong. You introspect on any motivations you might have that could have led you to want to conclude p. You revisit the evidence and reaffirm its actual bearing on p, etc. The problem is that this is highly unlikely to be an effective method of discovering and correcting for the impact of motivated reasoning. Even if your initial reasoning was biased, introspection is unlikely to reveal it even if you actively search for it. And, having failed to detect bias in your reasoning, you might wrongly conclude that your original reasoning was indeed strong and unbiased. Furthermore, your having made every attempt to eliminate bias as an explanation of your belief that p might lead you to become more confident than you were initially that p is correct. In this way, biased belief is often made even more biased by attempts to be objective in one's reasoning (Kenyon, 2014; Lord, Lepper, & Preston, 1984; McPherson Frantz & Janoff-Bulman, 2000). So while it may be perfectly reasonable to downgrade the epistemic position of someone on the other side of the political aisle by attributing motivated reasoning to them, it is not legitimate to absolve oneself of the same downgrading by reference to introspection.

This response to the debunking argument also presents a problem for a potential solution to the conflict between ABILITY and ACCURACY. Recall that my reason for downgrading your ACCURACY in BRILLIANT PARTISAN was roughly that the normal correlation between your ABILITY and my subjective probability of your being correct is reversed in the case of politically divisive propositions. But without such a positive correlation, there seems to be no motivation to adjust my view on the basis of your ABILITY at all (King, 2012). Why put any particular stock in the disagreement of a genius if you expect the genius to be mistaken about the kinds of matters under dispute? Absent a motivation to think that ABILITY correlates with ACCURACY, we should ignore ABILITY altogether and assess epistemic credentials on the basis of ACCURACY.

While this solution certainly has some initial draw, it leads to an odd consequence. Plausibly, if I assess your epistemic credentials by reference to your ACCURACY, I should do the same for my own. But we have seen that I have no reason to think that I am less likely to be subject to motivated reasoning than you are. Therefore, I should apply the same reasoning that I did with respect to your ACCURACY to myself. My own ABILITY has plausibly allowed my doxastic attitude to be less constrained by the actual force of the evidence than it would have been had I been less able. The consequence of that seems to be that I should think it less likely

that I am right the more familiar I am with the evidence, and the higher I rate myself with respect to the other aspects of ABILITY. But this seems strange on its face, and would seem to imply that I should think that I maximize my rationality with respect to political beliefs by avoiding evidence and taking steps to hamper by own cognition.[33]

5 Clustering

Elga (2007) offers a different line of argument that might provide us with an escape hatch from considering political disagreements to be epistemically significant. He asks us to consider two friends, Ann and Beth, who are on opposite sides of the political spectrum: "Does Ann consider Beth a peer with respect to [the claim that abortion is morally permissible]? That is: setting aside her own reasoning about the abortion claim (and Beth's contrary view about it), does Ann think Beth would be just as likely as her to get things right? The answer is "no". For (let us suppose) Ann and Beth have discussed claims closely linked to the abortion claim. They have discussed, for example, whether human beings have souls, whether it is permissible to withhold treatment from certain terminally ill infants, and whether rights figure prominently in a correct ethical theory. By Ann's lights, Beth has reached wrong conclusions about most of these closely related questions. As a result, even setting aside her own reasoning about the abortion claim, Ann thinks it unlikely that Beth would be right in case the two of them disagree about abortion" (2007, p. 493).

So on Elga's view, another reason I might have to reject the significance of political disagreements is that the clustering of beliefs about politically disputed propositions gives me dispute-independent reason to denigrate your epistemic standing. When I know something about your political position, as I do in BRILLIANT PARTISAN and MEDIOCRE PARTISAN, or if I know what you believe about a particular politically disputed proposition p, then I can infer something about your likely position on other politically disputed propositions q, r, and s. For example, if I know that you are a liberal democrat, or if I know that you believe that humans are causing global warming, then I can make a qualified guess about what you believe about the death penalty, abortion, gun control, and marihuana legalization (Kahan, 2015; Kahan et al., 2007). And I can use my belief that you are wrong about these other, but related, disagreements,

[33] Although he emphasizes different psychological mechanisms than I have here, Levy (2006) has suggested that we sometimes should in fact avoid evidence when seeking out and evaluating evidence carries a risk of worsening our epistemic position.

to denigrate your epistemic position with respect to p without violating the Independence principle.

I am not convinced that this argument gets us off the hook either. Consider that our respective political positions can be taken as a common determinant of our beliefs about p, q, r, and s. Elga's argument can then be construed as the claim that I am justified in thinking that my political position is generally more conducive to ACCURACY than yours is. But I have no dispute-independent reason to think that this is the case. The evidence on the political symmetry of motivated reasoning suggests that we are equally likely to have misconstrued the evidence so as to favor our politically congenial conclusions for all these propositions. To insist otherwise seems to be an instance of bias blind spot. Neither is there any dispute-independent reason to think that I am so lucky that my political position just happens to generally motivate me to defend conclusions that (fortunately) are also those that are in fact best supported by the evidence.

6 Reasons to denigrate

It is not always the case that disagreements in the political domain are epistemically significant. Sometimes, we do have dispute-independent reason to denigrate the epistemic standing of our interlocutor with respect to particular politically divisive propositions. The fact that the vast majority of experts agree that humans are causing global warming, and that such a verdict correlates with level of expertise in climate science, is one such reason (Cook et al., 2013, 2016). This fact gives an independent reason for someone who believes that humans are causing global warming to denigrate the epistemic standing of their interlocutor with respect to that question.

But we are not always so lucky that there is an overwhelming expert consensus or other dispute-independent ways to settle an issue. And absent such reasons, the claim I have made here is that it rarely will be the case that one will be able to denigrate someone's epistemic credentials on the basis of any purported differences in the quality of reasoning that stem from political position. Even when your belief about p actually reflects what the best evidence supports, this is not necessarily due to your having been any less biased in your reasoning about the relevant evidence than your interlocutor. Consider for example that people who believe that humans are causing global warming are no more likely to answer basic questions about climate science correctly than are people who deny it, and are equally likely to misconstrue evidence as being more supportive of their desired conclusion with respect to global warming than it is (Kahan, 2013, 2015). So presumably the superior epistemic position of believers is not the result of a

better understanding of climate science or more objective assessments of the evidence. Neither is it due to a greater general tendency to defer to scientific expertise. Among both conservatives and liberals, science is one the, if not the, most trusted institutions in society.[34] The problem is that our trust in science is selective. Whether we judge someone as a true expert tends to depend on whether they expound views that are congenial to our political beliefs (Kahan et al., 2011). A believer in global warming is, for example, more likely to dispute the scientific experts about the safety of nuclear energy than a denier of global warming is (Peters & Slovic, 1996; Pew Research Center, 2015). So rather than generally better ability or objectivity, the believers' superior epistemic positions are probably rather due, at least in part, to epistemic luck (Pritchard, 2005). They are fortunate enough that in this particular case, their political ideology inclines them toward the factual belief that happens to reflect what the evidence really supports.

7 Concluding remarks

Disagreements about politically divisive propositions are, then, often more epistemically significant than they have sometimes been perceived to be within the epistemology of disagreement. This leaves us with the unresolved puzzle of whether to ascertain the epistemic significance of such disagreements on the basis of ABILITY or ACCURACY. I have no hard answer to give – it seems to me to be a genuine puzzle with unwelcome consequences whichever route we go.

Instead, I close by suggesting that what holds for politics may hold for other divisive domains that plausibly evoke motivated reasoning, such as morality, and (for philosophers) philosophy. While I know of no direct evidence on this, it seems a distinct possibility that a similar empirical pattern might hold here – that as ABILITY increases; so does doxastic polarization within these domains. If so, disputants in these domains face the same problem that perceived ABILITY and ACCURACY are inversely correlated, and the task of adjudicating between these two in assessing the epistemic significance of disagreement.

[34] http://www.culturalcognition.net/blog/2014/11/25/conservatives-lose-faith-in-science-over-last-40-years-where.html

Article 3: Disagreement and the division of epistemic labor

Coauthored by Klemens Kappel

1 Introduction

Human agents routinely distribute epistemic labor between them. This is most salient when considering how many of one's beliefs are held due to their truth being testified by other people. Although we ourselves may lack direct evidence for the truth of such beliefs, believing them can nevertheless be rational, justified, or amount to knowledge, insofar as someone else has done the relevant epistemic labor required for such (Goldberg, 2010; Hardwig, 1985; Lackey, 2008a). The division of epistemic labor exemplifies the extent to which we depend on one another epistemically (Goldberg, 2011). Consider a standard case of division of epistemic labor: S_1 and S_2 are trying to find out whether p, which they know to be entailed by $q \wedge r$. Instead of S_1 and S_2 both expending the time and energy necessary to finding out whether q and whether r, S_1 works only on finding out whether q, while S_2 works on finding out whether r. If S_1 comes to know q, and S_2 comes to know r, and they each reliably testify their knowledge to the other, then both come to know p. In this case, S_1 epistemically depends on S_2 for his or her knowledge that q, and by extension that p.

In this article, we want to highlight another widespread way that humans divide their epistemic labor and thereby depend on one another epistemically. Consider again two subjects, S_1 and S_2, who are trying to find out whether p. There are reasons for and reasons against believing p, and these are of varying strength. Instead of both subjects looking for all the reasons both for and against p, and reflecting on their merits, they divide the epistemic labor such that S_1 looks for reasons to believe p, while S_2 looks for reasons to believe not-p. They then exchange their reasons, with S_1 looking for weaknesses in the reasons for believing not-p that S_2 presents, and vice versa. Suppose that over the course of their exchange, they manage to sort out the good reasons from the bad, and arrive at a good epistemic position with respect to p – one that is better than what they could have achieved with the same resources without dividing the epistemic labor. Here, S_1 and S_2 depend on one another not by mutual reliance on testimony, but by

dividing the search for and critical scrutiny of reasons. Call this a deliberative division of epistemic labor, to distinguish it from a testimonial division of epistemic labor.

There are features of human psychology that facilitate the occurrence of deliberative divisions of epistemic labor among agents who disagree. It has been amply demonstrated that our reasoning does not always serve a dispassionate search for truth. Rather, it is often directionally motivated: that is, it works to reach a construal of evidence that yields a conclusion that we already believe or that we want to justify (Kahan, 2016; Kunda, 1990; Nickerson, 1998; Taber & Lodge, 2006). In spite of its many pernicious epistemic effects, motivated reasoning can facilitate a deliberative division of epistemic labor between people who disagree. Motivated reasoning can increase one's ability to find good reasons in favor of one's view, and to critique reasons against it, compared to more dispassionate reasoning. When both sides of an issue are represented in deliberation, this results in a broad range of reasons being considered for each side, and the selective retention of the best ones. As a result, members of deliberating groups composed of people who disagree tend to arrive at beliefs that are better supported by the evidence than individuals reasoning alone, or members of groups composed of people who agree prior to deliberation (Laughlin & Ellis, 1986; Moshman & Geil, 1998; Schulz-Hardt et al., 2006; Trouche et al., 2014; van Knippenberg & Schippers, 2007).

While relying on the deliberative division of epistemic labor to improve our epistemic positions might seem like a good idea, doing so in an epistemically rational manner faces a challenge from the epistemology of disagreement. Here, it is commonly held that rationality requires a convergence of views upon the discovery of disagreement. But a convergence of views, we will argue, impedes the deliberative division of epistemic labor and its resulting benefits. This is the central problem that we wish to address.

Our aim is to argue that the benefits that accrue from the deliberative division of epistemic labor provide epistemic reason to maintain belief in the face of disagreement. We set the stage by presenting empirical evidence on the deliberative division of epistemic labor (section 2). We then address the problem from the epistemology of disagreement (section 3), and argue, mainly through a defense of a version of epistemic teleology, that the benefits that arise from the deliberative division of epistemic labor can make it rational to maintain belief in the face of disagreement (section 4). Finally, we discuss the significance of the deliberative division of epistemic labor for the notion of epistemic dependence, arguing that it suggests the need for an expansion of Goldberg's (2010) Extendedness Hypothesis (section 5).

2 Deliberative division of epistemic labor

The features of human cognition that facilitate the division of epistemic labor are those that we will place under the umbrella term of motivated reasoning.[35] As we will use the term here, motivated reasoning denotes the tendency to seek out, evaluate, and remember evidence in manners that are biased in favor of prior beliefs, or in favor of conclusions that are otherwise desired (Hennes et al., 2016; Jones & Sugden, 2001; Kunda, 1990; Nickerson, 1998). In cases where people are given the choice between viewing evidence that confirms their prior belief or evidence that disconfirms it, they tend to select confirmatory evidence (Frimer et al., 2017; Jones & Sugden, 2001; Taber & Lodge, 2006). When evaluating evidence, subjects tend to devote few cognitive resources to the scrutiny of evidence that supports their prior belief, and generally tend to rate such evidence as strong. Likewise, if an immediate cognitive response about the solution to a reasoning task supports a subject's prior belief, then deliberate reasoning is unlikely to be activated to check and possibly overwrite this response. On the other hand, evidence counting against prior beliefs tends to be heavily scrutinized. More time and effort is spent looking for defeaters, and this search may yield objections that, rightly or not, are taken to justify rejecting such evidence. Similarly, immediate cognitive responses that run counter to a prior belief will tend to activate deliberate reasoning in an attempt to find an alternative response (Dawson et al., 2002; Kahan, Peters, et al., 2017; Taber & Lodge, 2006). Subjects are more likely to recall evidence supporting their prior belief and to forget evidence against it, and are likely to misremember evidence as more supportive of their prior belief than it was (Hennes et al., 2016).

Unsurprisingly, motivated reasoning can result in many epistemic maladies. For example, it is implicated in belief perseverance, a tendency to maintain or sometimes even bolster belief in the face of undermining or rebutting defeat (C. A. Anderson, Lepper, & Ross, 1980; Nyhan & Reifler, 2010). It can lead one to entrench and rationalize factual errors, such as drawing a demonstrably false logical or mathematical conclusion (Evans, Barston, & Pollard, 1983; Kahan, Peters, et al., 2017). In social settings, it is implicated in 'groupthink' and extremism (Sunstein, 2002b).

Nevertheless, motivated reasoning can also carry benefits relative to more dispassionate cognition that aims only at accuracy. For example, when an intuitively appealing solution to a

[35] This umbrella includes the concepts of confirmation- or myside bias (Nickerson, 1998), biased assimilation (Lord et al., 1979), and motivated reasoning itself (Kunda, 1990).

task is wrong, a motivation to disbelieve the conclusion can increase the likelihood that one engages in the reasoning necessary to find the correct solution (Dawson et al., 2002; Kahan, Peters, et al., 2017). In such cases, experimental manipulations that decrease the magnitude of motivated reasoning tend to decrease performance (Munro & Stansbury, 2009).

In deliberative settings where group members disagree, the occasional epistemic bright side to motivated reasoning is amplified. Proponents of a view will tend to search selectively and efficiently for arguments in favor of that view. Opponents will tend to be highly scrutinizing of such arguments, and expend considerable cognitive resources on finding any flaws. Still, critical opponents nevertheless tend to be sensitive to argument strength, and are able to change their minds in response to strong arguments (Mercier & Sperber, 2011; Trouche et al., 2016, 2014; Vinokur & Burnstein, 1978).

There is ample evidence that such settings are epistemically fruitful. A striking example comes from studies using the Wason Selection Task, a test of conditional reasoning and the most widely used task in the psychology of reasoning (Wason, 1968).[36] Moshman & Geil (1998) had subjects attempt the task in groups of five or six, sometimes after first attempting the task alone. While only 9% of participants in an individual control condition gave the correct response, a rather typical result (Evans, 2002), 70% of participants who discussed the task in a group, and 80% who first attempted the task individually before discussing in a group, did so. These rather dramatic beneficial effects did not appear to be caused by individuals discovering the correct answer alone and then simply explaining it to their group. Even in groups where all members had arrived at an incorrect response prior to discussion, there were clear benefits to deliberation (indeed, all such groups found the right answer). This study did not directly contrast groups with or without disagreement: In no group was any answer unanimous prior to deliberation. Nevertheless, 100% of groups in which each member endorsed a different response found the correct solution, while this was "only" 75% of groups in which at least two group members endorsed the same response. Furthermore, transcripts from the group discussions, and follow-up experiments utilizing different tasks (Trouche et al., 2014), show that the benefits are due to a collective process of back-and-forth argumentation. The deliberative setting induces members to

[36] In the classical version of the task, participants are presented with four cards stipulated to have a letter on one side, and a number on the other. The visible sides of the cards might read E, K, 4, and 7. They are then asked which cards they must turn over in order to test the rule "if there is a vowel on one side, then there is an even number on the other side". The correct response to the task is taken to be the E card and the 7 card.

attempt to defend their own inferences and rebut the arguments others present in favor of selecting or not selecting a given card. Since people are in fact able to recognize when an inference of theirs has been shown by others to be invalid, and when another's is valid, invalid inferences are generally rejected while valid ones are retained over the course of deliberation.

Another example that disagreement can benefit group deliberation comes from experiments on 'hidden profiles'. In this type of experiment, a group has to find the correct solution to a problem (i.e. choosing the best of a set of n alternatives) after deliberating on the basis of evidence. All members of the group share some of the evidence prior to discussion. Each member also has unshared evidence available. In a hidden profile case, finding the correct solution requires aggregating all the unshared evidence during discussion, but groups typically fail to do this, focusing instead on the (misleading) shared evidence (Stasser & Titus, 1985). Schulz-Hardt, Brodbeck, Mojzisch, Kerschreiter, & Frey (2006) had groups of three members attempt to solve such problems, and manipulated the distribution of pre-discussion preferences (i.e. whether the total evidence held by each individual favored the same or different solutions, out of four possible solutions). In cases where each individual favored the same (wrong) solution prior to discussion, the hidden profile was solved only 7% of the time. This solution rate increased to 25% when one member favored a different response from the other two, but no one favored the correct solution, and to 28% when all three members differed in their preferred response but no one preferred the correct solution. In cases where one member favored the correct solution prior to discussion, the hidden profile was solved by 59% of groups when the other two members were in agreement, and by 65% of groups when all group members favored different responses.

Further evidence comes from studies on the effects of disagreement on work group performance and collective intelligence. Intellectual or informational diversity predicts success on a wide variety of intellective or creative tasks, where the group is tasked with deliberation about what the optimal solution is, rather than implementation, and this effect is mediated by disagreement about the task at hand (Jehn, Northcraft, & Neale, 1999; van Knippenberg & Schippers, 2007; Woolley et al., 2015).

Finally, although one should proceed with caution before applying results from purely theoretical models to real world deliberation, the notion that disagreement improves the outcomes of collective deliberation is also given some support from theoretical models of problem-solving, where a common result is that 'diversity trumps ability'. Diverse groups of agents are often

found to do better than more homogeneous groups, even if the homogenous groups consist of agents that are individually more competent (Hong & Page, 2004; Krause, James, Faria, Ruxton, & Krause, 2011; Mann & Helbing, 2016).

2.1 Some caveats

An important requirement for the deliberative division of epistemic labor to obtain is that the disagreement is with respect to whether some target proposition is true or false. Not all disagreements are of this nature. For example, agents can disagree about exact credences while leaning in the same direction (e.g. S_1 has credence 0.7 while S_2 has credence 0.9). If so, motivated reasoning will not support a division of epistemic labor, at least not to the same extent. Rather, the arguments brought to bear will tend to point in the same direction, and there will be less or no tendency for deliberators to look for flaws in each other's arguments. Group members will tend to be furnished with even more reasons that they are correct than they would if they reasoned in isolation, even if the proposition is actually false. Thus, the likely result would be what is known as group polarization (Vinokur & Burnstein, 1978a; Moscovici & Zavalloni, 1969; Sunstein, 2002). Another requirement is that there be a possibility of actual deliberation, i.e. the exchange of reasons. Exceptions to this include cases where the disagreement is purely due to different private evidence, e.g. of a perceptual nature, or cases where deliberation will simply not occur, e.g. for practical reasons. Certain types of issues and cognitive tasks may also yield greater benefits than others. So far, the biggest benefits have been found for somewhat artificial tasks with a demonstrably true solution, such as the Wason Selection Task, although this may partly be a methodological artifact: a demonstrably true conclusion is necessary to confidently say that there has been epistemic benefit (Kerr & Tindale, 2004; Laughlin & Ellis, 1986; Moshman & Geil, 1998; Schulz-Hardt et al., 2006). In deliberation about issues without a demonstrably true conclusion, there nevertheless does appear to be benefit and sensitivity to argument strength, also when the topic is moral or aesthetic rather than factual (Mercier, 2011; Vinokur & Burnstein, 1978). For deliberation about divisive political issues, most, but not all, studies find evidence of epistemic benefit (Barberá, 2015; Luskin, O'Flynn, Fishkin, & Russell, 2012; M. E. Price, 2012; Shih, Scheufele, & Brossard, 2013). We intend what follows to only apply to those cases where disagreement improves the quality of collective deliberation.

3 The challenge from the epistemology of disagreement

Given the above benefits, it is (at least for us) a natural thought that subjects would be rational in engaging in deliberation with those with whom they disagree in order to improve their epistemic position. However, there is, at least at first glance, a tension between realizing the benefits of the deliberative division of epistemic labor, and what rationality requires in cases of disagreement according to extant views in epistemology: a reduction in the magnitude of the disagreement upon its discovery.

We can formulate the (apparent) tension as follows:

> (1) The deliberative division of epistemic labor is promoted by subjects' motivated reasoning in defense of their belief.
>
> (2) Withholding judgment or reducing confidence in a belief decreases subsequent motivated reasoning in defense of that belief during deliberation.
>
> (3) On extant views in the epistemology of disagreement, rationality requires that one withhold judgment or reduce confidence upon the discovery of disagreement.
>
> (4) Conforming to rationality diminishes the deliberative division of epistemic labor.

Finding out whether this apparent tension is real upon closer inspection requires some work. Let us start with (1). A natural objection is that subjects can choose to divide the epistemic labor, without the need for motivated reasoning. It may not be necessary that the subjects actually believe what they defend; only that they accept the proposition, or play devil's advocate, for the purposes of promoting the division of epistemic labor. However, although there is evidence that such artificial disagreement is better than no disagreement at revealing unshared information, it is not as efficient at doing so as genuine disagreement. More importantly, artificial disagreement, unlike genuine disagreement, does not appear to carry any benefits to the quality of the outcome of deliberation (Greitemeyer, Schulz-Hardt, Brodbeck, & Frey, 2006; Schulz-Hardt et al., 2002). A plausible explanation for this result is that arguing for a proposition that one does not actually believe fails to activate (the right kind of) reasoning to the same extent as arguing for a proposition one does believe (Mercier & Sperber, 2011). This explanation is bolstered by studies on individuals where manipulations that decrease motivated reasoning can lead to decreases in performance (Munro & Stansbury, 2009).

This brings us to (2). It seems possible that someone who suspends judgment or reduces confidence, for instance on the basis of higher order evidence constituted by the fact of

disagreement, might nevertheless (at least temporarily) retain all the first order evidence for his or her belief and be inclined to reason about the proposition in the same manner during deliberation.[37] If so, reducing confidence in response to disagreement might not have any untoward consequences for the division of epistemic labor. While it is hard to experimentally manipulate belief and observe its effects on reasoning, there is indirect evidence that a change in doxastic state can have immediate impact on one's attitude toward one's first-order evidence and on one's subsequent reasoning. For example, in Asch's (1956) classical conformity studies, subjects who go along with an obviously mistaken majority response to a perceptual task do so in part because the social circumstances lead them to experience uncertainty about the veracity of their perceptual evidence (Abrams et al., 1990). This suggests that a reduction of confidence due to disagreement sometimes leads to doubt about one's first-order evidence, even when the evidence is in fact highly probative. With respect to effects on reasoning, an example comes from experiments on 'choice blindness'. In these experiments, subjects are asked to report their attitude toward some proposition. For example, they may be asked to indicate their agreement, on a numerical scale, with a series of claims such as "Gasoline taxes should be lowered". After reporting their attitude toward a claim, they are presented with their response and invited to explain it. However, on some trials, the experimenters use sleight-of-hand to present subjects with the opposite attitude of what they indicated and invite them to explain this response instead.[38] So, a subject who had indicated high agreement with "Gasoline taxes should be lowered" would now be presented, as if it were their own, with a response indicating high agreement with the claim "Gasoline taxes should be raised" and asked to explain why they hold this attitude. Only on quite few trials do subjects detect the mismatch between their original response and what they are presented with. Moreover, when the mismatch is not detected, and subjects accept the presented attitude as their own, they tend to rationalize the attitude they are presented with. They display the hallmarks of motivated reasoning in defense of a proposition that, a few seconds ago, they indicated disbelieving.[39] Thus, a change in what one believes (or, perhaps more accurately, what one believes that one believes) can lead to a complete shift in the direction of bias in one's subsequent reasoning. Because the experimenters only changed

[37] We thank Giacomo Melis and an anonymous reviewer for pressing us about this issue.

[38] See (Hall et al., 2013; Hall, Johansson, & Strandberg, 2012; Johansson, Hall, Sikström, & Olsson, 2005; Rieznik et al., 2017; Trouche et al., 2016) for details.

[39] See the supplemental materials to Hall et al. (2012) for some striking examples of such justifications in the moral domain.

responses to the negation of an indicated belief, we do not have direct evidence speaking to the effect of changing a report of belief to a report of withholding judgment, or to the same belief held with less confidence. In light of the dramatic effects of the reversal manipulation, however, we find it plausible that there would be an effect of decreasing confidence as well, especially since motivated reasoning is typically detected mainly in subjects who hold strong beliefs (Taber & Lodge, 2006).

Assessing (3) requires a more thorough consideration of what epistemologists have taken rationality to require in cases of disagreement. After all, if a plausible view of disagreement holds that there is no rational pressure for disagreement to diminish upon its discovery, there is no tension to resolve. Most discussion has focused on cases of the following kind: Suppose that two people, who consider each other to be roughly epistemic peers[40] on a topic, discover that they disagree. What, if any, belief change is rationally required?

On conciliatory views of disagreement, the rational response for both parties to such a disagreement is to become less confident in their belief. Given that the parties consider each other to be peers, neither is in a position to justifiably think it more likely that their interlocutor rather than themselves have made a mistake or is less accurate (Christensen, 2014a; Elga, 2007; Matheson, 2009). Alternately, conciliatory views have been argued for by showing that under certain conditions, conciliation increases accuracy as measured by scoring rules such as the Brier Score (Kopec, 2012; Lam, 2011, 2013). This brief statement naturally hides many variations. At the most conciliatory end of the spectrum, one finds the view that one should give equal weight to one's own view and that of an epistemic peer, and 'split the difference' between the parties' initial credences (Elga, 2007). Other views require less dramatic reductions of confidence while remaining in the conciliatory end of the spectrum. We can set aside these and other complications: What matters for our purposes is the shared view that rationality requires a convergence of views upon discovering disagreement.

Previous work on epistemic benefits of disagreement (Dunn, 2013; Matheson, 2014; Moffett, 2007) has focused on how such benefits may pose a problem specifically for

[40] Epistemic peerhood is defined in various ways. One definition holds that agents are peers if there is cognitive and evidential equality between them (Lackey, 2008b), or if they are approximate equals with respect to various intellectual virtues such as intelligence or diligence (Christensen, 2009). Another holds that my peer is someone I consider equally likely as myself to be correct, conditional on our disagreeing, prior to discovering disagreement (Elga, 2007).

conciliatory views. We think the tension is broader than this. The deliberative division of epistemic labor depends on the presence of opposing beliefs about a target proposition, and is in tension with any view whose requirements threaten this precondition. It is relatively clear how conciliatory views might do so. But we suggest that so do several views that lie relatively close to the opposite end of the spectrum, typically denoted as 'steadfast'. A steadfast view is one according to which at least one party to a peer disagreement can be rational in maintaining confidence (Enoch, 2010; Kelly, 2013; Titelbaum, 2013; Weatherson, 2016). On the extreme end of the spectrum are views where neither party in a case of peer disagreement is required to reduce confidence. This includes views that emphasize that the direct access one has to one's own reasons and evidence in virtue of the first person perspective can make it justified to remain unmoved. Since both parties to a disagreement have a first-person perspective, both can remain unmoved (Wedgwood, 2007). The verdicts of such views are not in tension with a deliberative division of labor. However, less extreme steadfast views typically do not imply that the there should continue to be opposing beliefs after the discovery of disagreement. Several steadfast views hold that there is an asymmetry that allows one person, namely the one who has the true belief, the more justified or rational belief, or knowledge, to be less moved by the disagreement than conciliatory views would have it. This may, but need not, mean that the person who 'got it right' should remain wholly unmoved. However, similar considerations do not apply to the person who did not get it right. For this person, responding to both the original evidence and the additional evidence acquired by discovering disagreement requires a move towards the view of the other party. As such, steadfast views too can require a decrease in the magnitude of disagreement. For example, consider the Knowledge Disagreement Norm (Hawthorne & Srinivasan, 2013), which holds that subjects ought to suspend judgment if they cannot retain knowledge by remaining steadfast or attain knowledge by adopting their interlocutor's belief. This norm is certainly steadfast in that it warrants a party to a peer disagreement to remain completely unmoved insofar as this allows them to retain knowledge. Nevertheless, there is no circumstance where applying the norm does not result in a diminished disagreement. If one party can retain knowledge of the disputed proposition p, the other cannot possibly retain knowledge of not-p. So either 1) both parties come to agree by knowing p, 2) they come to agree by suspending judgment, or 3) one person retains knowledge that p while the other suspends judgment regarding p.

While both conciliatory views and (most) steadfast views therefore seem to be in conflict with the deliberative division of epistemic labor, a natural objection to (3) concerns the scope of the

relevant epistemic principles. It might be thought that when epistemologists describe what rationality requires in cases of disagreement, what they have in mind are relatively idealized cases that are importantly different from what one finds in psychological experiments and more ordinary cases of disagreement. If (3) only applies to these idealized cases, and the division of epistemic labor occurs in non-idealized cases, then (4) is false. More specifically, it might be that (3) only applies in cases that feature epistemic peerhood and full disclosure, which will seldom (if ever) be the case in ordinary disagreements.

Our view is the following: Epistemic peerhood is doubtlessly useful as a theoretical tool for understanding the epistemic significance of disagreement as such, absent any concerns about asymmetries in evidence, ability, or accuracy. However, it is quite rare that two people share exactly the same (or at least equally good) evidence, are exactly equally likely to get the answer to a class of questions right, or are equals with respect to epistemic virtues. It might even be the case that we should rarely believe of any disagreement that it is a peer disagreement, if we realize that such cases are exceedingly rare (King, 2012). While we acknowledge that many of the real life cases in which the division of epistemic labor takes place will fall short of epistemic peerhood, we do not believe this has any consequences for whether (3) applies to such cases. Ordinary disagreements are still epistemically significant (Matheson, 2015). For one, it may be highly uncertain who is in a better or worse epistemic position, and this uncertainty can raise doubts about the rational status of our beliefs (King, 2012). Second, even if there is known asymmetry, disagreement can apply rational pressure. Suppose that two friends disagree about a proposition. Based on their track record, they know that when they have disagreed about similar questions in the past, one of them has been wrong twice as often as the other. Nevertheless, learning about the disagreement should have epistemic significance even for the friend who has tended to be right most of the time. After all, conditional on their disagreement, there is a one in three chance that this person is wrong; enough we think to raise some doubts about the correctness of their reasoning and to make ignoring the disagreement altogether rationally suspect (Elga, 2007). So, while the epistemological literature has focused on cases featuring epistemic peerhood, there is no reason to think that (3) only applies to such cases.

The notion of full disclosure might seem to pose a greater problem for the claim that (3) holds in cases with the potential for a division of epistemic labor. In the literature on disagreement, the interlocutors are often assumed to have shared all their available evidence. But of course, sharing all evidence, if this includes collective reasoning about the evidence, amounts to instantiating a division of epistemic labor. If (3) applies only after full disclosure, then it seems that the tension

dissipates, as any benefits will already have been obtained when the requirement to reduce confidence applies. However, we do not think the tension is so easily resolved. As we see it, full disclosure is, like peerhood, mainly a useful theoretical tool for understanding the epistemic significance of disagreement without the polluting influence of concerns about asymmetries with respect to evidence. If we do not know that we share evidence, then disagreement may not to the same extent provide evidence that one of us has made a mistake, since a difference in views may simply reflect proper appreciation of different bodies of evidence. So for the purposes of investigating the impact of a certain type of higher-order evidence about the rationality of one's beliefs, full disclosure is a useful construct (Christensen, 2010). But this doesn't mean that disagreements lacking full disclosure lack normative impact. Rather, in such disagreements there is *both* the higher order evidence that one may have erred, and evidence that one's interlocutor has pertinent evidence that one lacks. Both of these possibilities are plausibly reasons to moderate one's view.

So, on closer inspection we find that there is a genuine, and not merely apparent, tension between what extant views in the epistemology of disagreement take rationality to require, and attaining the epistemic benefits that arise from the division of epistemic labor.

4 Epistemic rationality and the division of epistemic labor

In this section, we defend the claim that the benefits that arise from the division of epistemic labor can make it all things considered epistemically rational to maintain belief in the face of disagreement, even though, as the epistemology of disagreement has shown, there are also (pro tanto) reasons to reduce confidence.

A motivation for this claim comes from considering the nature of inquiry. Suppose that inquiry about p consists in trying to find out whether p (Kelp, 2014).[41] Ordinarily, this goal is best served by believing what one's evidence about p supports. But the evidence we have presented suggests that some cases of disagreement are unusual insofar as responding to a piece of evidence pertinent to whether p (i.e. the fact of disagreement) is, temporarily, not the best way of finding out whether p, because responding to this evidence prevents obtaining the benefits to

[41] This can be cashed out in various ways: Coming to know whether p, coming to have a true belief about whether p, coming to have a justified belief about whether p, or coming to have a justified true belief about whether p.

finding out whether p that arise from a division of epistemic labor. Insofar as epistemic rationality serves inquiry, it seems odd to say that rationality requires responding to this evidence, even when doing so means hindering inquiry.

However, supporting this judgment by reference to common theories of epistemic rationality is difficult. The main difficulty is that it seems to require assessing the belief a subject holds upon discovery of disagreement by looking at its conduciveness to epistemic benefits. But epistemic assessment by common theories of rationality does not involve any consequences of holding a belief. Take a standard version of evidentialism, according to which the epistemic assessment of S's doxastic attitude toward p at time t is determined solely by S's total evidence at t. If we assume, in line with extant views in the epistemology of disagreement, that disagreement constitutes evidence in support of reduced confidence, then it is clear that evidentialism would require reduced confidence regardless of any benefits to maintaining confidence. Next consider process reliabilism. According to process reliabilism, the epistemic status of a belief is determined by whether it was formed through a reliable process. It does not matter whether the belief will have good downstream consequences, or constitute an input to a reliable process. So although the process of collective deliberation with someone with whom one disagrees is itself a reliable belief-forming process, this does not feature in the assessment of belief upon the discovery of disagreement, since this process did not produce the belief (Goldman, 2015). So neither evidentialism nor process reliabilism will view the benefits of the division of epistemic labor as relevant to epistemic assessment.

We think this leads to some awkward conclusions for traditional modes of epistemic assessment in these cases. To help us discuss these issues, we'll draw on the following case:

> S_1 believes that p at t_1, and anticipates meeting S_2 in at t_2. S_1 anticipates that the question whether p will come up. At t_1, S_1 considers the possibility that S_2 will disagree about p. If it turns out that S_2 disagrees, S_1 knows that a reduction of confidence at t_2 would be a more rational response to her evidence at t_2 than remaining confident. But she also knows that if she reduces confidence at t_2, then her doxastic attitude toward p at t_3, after they have deliberated, will be a worse reflection of her evidence than if she does not reduce confidence at t_2.

What should S_1, at t_1, think about what it would be rational for her to believe at t_2? Let's focus on evidentialism, and suppose that the answer is "whatever my evidence supports at t_2". This would

mean that rationality with respect to p is self-defeating in a certain way. S_1 at t_1 could anticipate being less rational at t_3, and indefinitely ahead in time, if she is rational at t_2. It would therefore seem natural for her at t_1 to anticipate herself regretting, at t_3, letting her belief be governed rationally at t_2. And at t_3, she would know that she is now less rational than she could have been, and as such it would be natural for her to regret having been rational at t_2.[42] Since S_1 had all the considerations that led to the regrettable state of affairs at t_3 available to her at t_1 and t_2, it could be argued that she rationally *should* regret reducing confidence (McQueen, 2017). Now, there can clearly be beliefs that one can anticipate regretting but are nevertheless epistemically rational. This is the case if the reason to regret the belief is frustration of one's practical goals, for example. It is less clear what the rational import is when one can anticipate purely epistemic regret, i.e. regret brought about by frustrated epistemic rationality. On the one hand, one might take such regret to be unfortunate, but to not have any consequences for the rationality of the regrettable belief. On the other hand, it might be thought that part of being epistemically rational is that one avoids taking steps in one's inquiry that one can anticipate regretting purely on the grounds of costs to the goals of said inquiry. In other words, perhaps one should avoid steep temporal discounting of the goal of inquiry.

There is practically no discussion of this kind of epistemic regret in the literature that we know of. However, the issue of regret is prominent in the literature on dynamic choice, which the above case is an epistemic version of (Andreou, 2017). In a typical dynamic choice problem, an agent has preferences about a future outcome that are best served by his or her resisting a future temptation. While the agent therefore prefers and intends to resist the temptation, he or she anticipates that upon encountering the temptation, his or her preference will shift in favor of the temptation. The question is whether there is an account of rationality that allows the agent to stick to his or her prior intention. Bratman (1999, 2012) has argued for a planning conception of rationality that includes a "no-regret" condition. One ought to care about how one will see things at the conclusion of one's plan and avoid actions and adjustments to one's plan that one will regret in the future. One is rational in resisting the temptation in part because the anticipated regret of doing otherwise, from one's standpoint upon encountering the temptation, provides a reason to change one's current preferences (Bratman, 2014).

[42] An anonymous reviewer pointed us toward a discussion by White (2010) on cases that are somewhat similar. White argues that one need not feel any epistemic discomfort in cases where one knows that it would seem as if one is rational regardless of whether one is in a good case (where one actually is rational), or a bad case (where one is not), even if one has no independent way to know that one is in a good rather than a bad case. But note that in our case subjects do have an independent reason to think that they are in a "bad case" at t_3.

Transposing this line of reasoning to the epistemic case, one might take the relevant plan to be inquiry with respect to p. The goal of inquiry would be best served by not responding to the evidence constituted by the fact of disagreement at t_2. So S_1 might at t_1 form an intention to retain belief at t_2. However, she can anticipate that she would be rational in responding to the evidence once she encounters S_2. Could there be a reason for her to stick to her intention? On a planning conception of epistemic rationality, from S_1's perspective at t_2, the anticipated epistemic regret of reducing confidence would amount to such a reason. However, it is admittedly not clear that the planning conception of rationality can be neatly transposed from the practical to the epistemic domain in this manner. There seems to be an important difference between anticipated regret making a difference to what one prefers, and anticipated epistemic regret making a difference to what one rationally ought to believe. So the problems of rational self-defeat and epistemic regret may therefore not suffice to show that it is rational to maintain belief upon encountering disagreement. We do however think they suggest that something is problematic about the opposite conclusion.

Another way of pressing a similar point is to consider that the possible version of S_1 at t_3 who did not reduce confidence at t_2 is an expert about p relative to the possible version of S_1 at t_3 who did reduce confidence at t_2, and relative to S_1 at t_1. That is, this possible version of S_1 at t_3 has a belief that is better supported by the evidence. If you know that someone is an expert in this sense, then you should defer to him or her with respect to p.[43] A difficulty here compared to the cases of expertise usually discussed is that S_1 does not know at t_1 exactly what doxastic attitude this version of herself at t_3 will have, only that it will be better supported by the evidence. So she cannot adopt the doxastic attitude that would amount to a direct deference to the expert. But she does know what doxastic attitude she should adopt at t_2 in order to make it the case that she will believe what the expert believes at t_3: she should maintain belief.

While we think the above considerations are suggestive, a more straightforward argument for the conclusion that maintaining belief in the face of disagreement is rational can be fielded by relying on normative epistemic teleology for the assessment of rationality. Roughly, normative epistemic teleology is a family of views whereby the epistemic assessment of a doxastic attitude is determined by the extent to which holding the attitude fosters attainment of some relevant set

[43] When that expert is your future self, this seems to be a guiding intuition behind van Fraassen's (1984) Reflection Principle, and of time-slice centric versions of that principle (Hedden, 2015).

of goals or ends (Ahlstrom-Vij & Dunn, 2014; Carr, 2017; Kopec, 2017).[44] To borrow Selim Berker's (2013a) terminology, a teleological theory will contain 1) a theory of final value that describes what states of affairs have value (or disvalue) as ends in themselves, 2) a theory of overall value that assigns rankings to entities (e.g. beliefs) based on their promotion of the final value, and 3) a deontic theory that assigns deontic properties (e.g. "rational" or "justified") to entities on the basis of the theory of overall value. It would be beyond the scope of this article to argue for epistemic teleology as a general approach here.[45] Rather, our more modest aim is to show that there are versions of teleology on which the benefits arising from the division of epistemic labor render belief in the face of disagreement rational, and that escape some prominent recent criticisms of teleology.

To illustrate how a teleological theory could countenance the benefits that arise from the division of epistemic labor in epistemic assessment, consider a very simple veritistic teleological theory (which we would not endorse), where believing truths has final value (and believing falsehoods has final disvalue), beliefs are ranked according to their promotion of truth (and avoidance of falsehoods), and beliefs are rational to the extent that they promote believing truths and prevent believing falsehoods. If, in a case of disagreement, maintaining belief improves the attainment of truth following deliberation relative to a reduction of confidence, maintaining belief upon the discovery of disagreement would be epistemically rational on such a theory.

Of course, epistemic teleology is not without critics, and we will have to engage with some of that criticism.

A common line of critique is that, unlike actions, doxastic attitudes are not the kind of thing we (can) decide upon on the basis of their promotion of some end. When we engage in conscious reasoning about whether to believe p, our reasoning is governed by evidence bearing on the question whether p, not the consequences of adopting this or that doxastic attitude. If, for instance, one consciously reasons that the evidence, including the fact of disagreement, supports

[44] We can distinguish normative epistemic teleology from meta-epistemic teleology, which holds that epistemic norms have force in virtue of promoting value. One could be a meta-epistemic teleologist while being holding a non-teleological view about normative epistemology, or vice versa. For example, one could hold that epistemic norms have force due to their promotion of one's practical goals while being an evidentialist about normative epistemology (Cowie, 2014).

[45] For recent defenses of epistemic teleology, see (Ahlstrom-Vij & Dunn, 2014; Klausen, 2009; Kopec, 2017; Talbot, 2014).

suspension of judgment in p, it is without a doubt psychologically difficult, perhaps impossible, to believe p even if one is aware that so doing would carry future benefits for one's ability to properly assess what the evidence supports (Kelly, 2002). Suppose for the purposes of argument that it is correct that we cannot consciously adopt beliefs on the basis of teleological reasoning. This would plausibly show that teleology falls short as an epistemic decision procedure: even if we were to know what we ought to believe according to a teleological norm, we might not be able to believe that which we know we ought to. But this does not show that teleology could not be a valid criterion of rightness. Beliefs that are not formed through conscious application of a norm can nevertheless be assessed according to the norm. Consider beliefs that are adopted due to the operation of some unconscious belief-forming mechanism. Such beliefs can be evaluated epistemically. We can assess whether they are supported by the subject's evidence, even if the subject has not formed the belief by consciously responding to that evidence (i.e. we can assess whether the beliefs are propositionally justified). Similarly, we can engage in teleological assessment of subjects' beliefs in cases of disagreement even if the belief is not adopted or maintained by explicit teleological reasoning.

A different line of criticism against teleology has been leveled by Berker (2013a, 2013b).[46] In his earlier work on the topic, Berker (2013a) argues that epistemic teleology fails because it runs afoul of what he calls the epistemic separateness of propositions, modeled on the ethical separateness of persons stressed by Rawls (1971). In ethics, murdering one person in order to prevent the murders of five others might be held to be wrong because the separateness of persons means that the value to the five cannot be traded off against the disvalue to the one. Berker argues that, similarly, promoting the epistemic value of one's doxastic attitudes toward p_2, p_3, …, p_n, cannot outweigh a cost to the epistemic value of one's doxastic attitude toward p_1 due to the separateness of propositions. And, similar to how consequentialist theories in ethics will tend to allow or mandate tradeoffs even in problematic cases due to their disregard for the separateness of persons, teleological theories in epistemology will tend to allow or mandate epistemic tradeoffs even in problematic cases due to their disregard for the separateness of propositions. Berker illustrates this with cases such as the following (Berker, 2013a, p. 364):[47]

"I am a scientist seeking to get a grant from a religious organization. Suppose, also, that I

[46] Other recent critiques of epistemic teleology include (Greaves, 2013; Jenkins, 2007; Littlejohn, 2012).

[47] Berker traces this type of critique to Firth (1981).

> am an atheist: I have thought long and hard about whether God exists and have eventually come to the conclusion that He does not. However, I realize that my only chance of receiving funding from the organization is to believe in the existence of God: they only give grants to believers, and I know I am such a bad liar that I won't be able to convince the organization's review board that I believe God exists unless I genuinely do. Finally, I know that, were I to receive the grant, I would use it to further my research, which would allow me to form a large number of new true beliefs and to revise a large number of previously held false beliefs about a variety of matters of great intellectual significance. Given these circumstances, should I form a belief that God exists? Would such a belief be epistemically rational, or reasonable, or justified?"

Berker holds it to be obvious that the belief would not be rational, and we will not dispute that judgment here. What is wrong with the scientist's belief that God exists, on the principle of the separateness of propositions, is that the epistemic value promoted by that belief accrues to other propositions, and so cannot outweigh the cost to the value of the belief about the existence of God. The separateness of propositions means, according to Berker, that such benefits are simply irrelevant to the epistemic assessment of the scientist's belief in God. But, Berker argues, some teleological views will tend to yield the opposite conclusion. In the above case, this is true for our simple veritistic norm from before. And while more sophisticated versions of teleology might be able to escape countenancing the tradeoff in the above case, Berker claims, and presents modified cases to show, that the underlying problem – the disregard for the separateness of propositions – means that it will always be possible to construct cases that pose a problem for a revised version of teleology. A teleological theory might be restricted specifically to avoid this problem, i.e. by holding that a belief with a certain propositional content only counts as promoting epistemic value if it promotes that value for beliefs with the same propositional content. While this would avoid violating the separateness of propositions, Berker claims that such amendments will run into further problems.

Specifically, Berker's (2013b) later work expands his criticism of teleology by presenting cases that are problematic for teleology although they do not involve any violation of the separateness of propositions. These cases involve beliefs that are self-fulfilling, and thereby causally promote their own epistemic value, but nevertheless seem irrational. In Berker's Jane Doe case, a woman who suffers from an illness can increase the odds that she will recover if she, against the evidence she has available, manages to adopt the belief that she will recover:

> "…Let us suppose that she is not aware of this, and, to fix on some numbers, let us suppose that she has a 10% of recovering if she does not believe she will recover and a 90% of recovering if she does believe she will recover (where, moreover, the relevant percentages are caused by her being in the relevant doxastic state, not merely correlated with her so being)" (Berker, 2013b, p. 376).

Berker's verdict is that it is obvious that it would not be epistemically rational for Jane Doe to believe that she will recover, even if she does in fact recover due to her holding this belief, which would then turn out to have been true all along. A plausible reason why is that rational belief has a world-to-mind direction of fit. Rational belief aims at reflecting the world, not at changing the world so as to bring it in line with belief.[48] But, as Berker argues, veritistic versions of teleology have difficulty with capturing this aspect of rationality and will struggle not to yield the verdict that the belief in question is rational, even if they contain provisions against allowing tradeoffs that violate the separateness of propositions. After all, the Jane Doe case is one where belief promotes truth, at no cost to truth, and does so without violating the separateness of propositions.

To answer these charges, defenders of teleology have largely responded by agreeing with the verdict that belief in the cases Berker presents is irrational, but rejecting that their particular views do in fact license any such irrational belief in those kinds of cases (Ahlstrom-Vij & Dunn, 2014; Goldman, 2015; Kopec, 2017).[49] We are however in a different dialectical position than these authors. Our point of departure is an explicit endorsement of maintaining belief in disagreement cases for teleological reasons, although such continued belief is not supported by one's evidence. Structurally this is akin to the instances of belief that Berker takes to be obvious counterexamples to teleology. So to get off the ground it seems we need to confront the possible objection that continued belief in disagreement cases is simply obviously irrational. To do this let's consider a Berker-style case of disagreement:

> Suppose that S_1 believes that p on the basis of a body of evidence e. Due to motivated reasoning, S_1 has construed e as more supportive of p than it is. S_1 discovers that S_2 believes that not-p, also on the basis of e. S_2's disagreement is evidence that S_1 is overconfident about p. However, if both engage in deliberation while maintaining their beliefs, S_1's resulting doxastic attitude toward p would be a better reflection of what e in

[48] We borrowed this formulation from an anonymous reviewer, with thanks.

[49] Although Ahlstrom-Vij and Dunn argue that the tradeoff in Berker's Prime Numbers case is permissible.

fact supports, compared to his or her doxastic attitude toward p if they deliberate after reducing confidence. Given these circumstances, should S_1 maintain belief in p? Would such a belief be epistemically rational, reasonable, or justified?

We do not think maintaining belief in this case would be obviously irrational to the extent that the question is settled. We suspect that this weaker intuition is partly due to the absence of major deviations from ordinary epistemic practice. Berker's scientist and Jane Doe believe what their evidence overwhelmingly suggests in untrue, whereas subjects in a case of disagreement 'merely' continue believing what their first-order evidence, by their lights, supports, in anticipation of deliberation that will hopefully disclose the reasons for the disagreement. Furthermore, Berker's scientist and Jane Doe engage in wishful thinking in order to form the value-promoting beliefs, and belief based on wishful thinking is typically a prime example of irrational belief. In contrast, it is (rightly or wrongly) common for human agents in cases of disagreement to maintain their confidence during their ordinary doxastic practice. While these are of course not reasons in themselves to countenance maintaining belief, it raises the question of whether the strong intuitions in cases like Berker's occur in part because of the element of wishful thinking, abject denial in the face of overwhelming evidence, or similar deviations from ordinary epistemic practice.

This is not to say that the prospect of maintaining belief in the disagreement case does not evoke a degree of epistemic unease. Maintaining confidence in the face of evidence that you're overconfident is, taken in isolation, surely not rational. But, similar to how even permissible tradeoffs can generate unease in ethical cases, we think such epistemic unease is to be expected and does not settle whether belief is rational. And, as we saw with the issue of epistemic regret, a purely evidentialist verdict can yield its own element of unease.

So let us proceed under the assumption that intuitions about the case alone are not sufficient to reach a verdict about what is rational, and turn to whether maintaining belief in the disagreement case falls into Berker's pitfalls of violating the separateness of propositions and the requirement that beliefs have a world-to-mind direction of fit.

Both the costs, in the form of a failure to respond to the fact of disagreement, and the benefits, in the form of an improved epistemic position following the division of epistemic labor, accrue to one's doxastic attitude toward p, rather than some other propositions. What generates a reason to maintain belief that p is exactly that this promotes reaching the verdict about p that is best supported by e. This might suggest that the case does not involve a violation of the separateness of propositions. However, in a footnote, Berker (2013a, p. 365, fn. 40) explicitly

states that inter-temporal, intra-propositional tradeoffs also violate the separateness of propositions:

> "More precisely, I should speak here of 'the epistemic separateness of propositions-at-a-time' since it is also epistemically irrelevant whether or not a belief in p at a given time conduces toward the promotion of true belief and the avoidance of false belief with regard to that same proposition at later times."[50]

Berker does not offer a reason for expanding the principle to include intra-propositional tradeoffs in this way, and the expanded principle is not illustrated with any purportedly problematic cases.[51] It is perhaps worth noting that the analogy with ethics also breaks down at this point, as there is no general principle of the ethical separateness of persons-at-a-time. For instance, it would be a mistake to say that a dentist who performs a treatment that causes the patient mild discomfort at t_1 in order to avoid much greater discomfort to the patient at t_2 is violating the ethical separateness of persons. We see no good reason why it is not a similar mistake to say that intra-propositional epistemic tradeoffs violate the separateness of propositions. So even if we should accept a principle about the epistemic separateness of propositions,[52] the disagreement case does not violate such a principle, and an expanded principle against intra-propositional tradeoffs seems unmotivated.

How does the disagreement case then fare with respect to observing the world-to-mind direction of fit? The disagreement case is not one where belief in the proposition causes it to be true. What it causes is an improved appreciation of what the evidence supports at the termination of deliberation, but the evidence, properly appreciated, may turn out to support whichever doxastic attitude toward the disputed proposition. So the reason one has to maintain belief in the face of disagreement is entirely due to facilitating an improved position from which to match one's doxastic attitude to the world.

---

[50] Berker is here targeting teleology with a veritistic theory of final value, but the same point would supposedly hold for the promotion of any other epistemic value.

[51] Theories of time-slice rationality, which hold that the relationship between two time-slices of the same person is not importantly different from the relationship between different persons, for the purposes of rational evaluation, might provide a possible way to motivate this view (Hedden, 2015; Moss, 2015). Engaging thoroughly with time-slice rationality would be beyond what we have space for here.

[52] Goldman (2015) offers some considerations against such a general principle.

The disagreement case therefore appears to escape the most prominent recent criticisms of teleology. However, there is a further challenge in showing that there are plausible teleological norms that are able to countenance belief in this case without thereby being forced to countenance the purportedly irrational instances of belief in cases such as those Berker presents. We think that there are such versions.

For example, consider a version of teleology that we, following Kopec (2017) will call evidential teleology. Evidential teleology holds that "a doxastic attitude generates final epistemic value to the extent that it accords with the possessor's total body of evidence. Furthermore, the evidential teleologist holds that a doxastic attitude is rational to the extent that it promotes the attainment of this kind of epistemic value" (Kopec, 2017, p. 19). In the disagreement case, evidential teleology would deem maintaining belief upon the discovery of disagreement as rational, given that after deliberation, the subject's doxastic attitude toward p more accurately reflects her total evidence if he or she maintains belief upon the discovery of disagreement than if he or she reduces confidence. So maintaining belief upon the discovery of disagreement is rational due to its promotion of final epistemic value. Note however that evidential teleology, as a (teleological) species of evidentialism, is also able to capture the feeling of unease that accompanies the verdict that maintaining belief is rational. After all, when one discovers disagreement, one thereby receives evidence to the effect that one is overconfident, and a failure to respond to this evidence constitutes a (temporary) reduction in the final epistemic value of the belief. We think this is a positive feature of the view. The case does, after all, involve a tradeoff, and any norm that is completely blind to the cost side of the equation is missing something important about the case.

So how does evidential teleology fare in dealing with Berker's key objections to teleology? With respect to the separateness of propositions, Kopec argues that, supposing that Berker's scientist starts out only having beliefs that are supported by his or her total evidence, it is not the case that he or she should believe that God exists. For in addition to the new true beliefs (which we can assume would also be supported by his or her evidence) the scientist would gain one belief that would not be supported by his or her total evidence, namely the belief that God exists. So the scientist would go from a state of affairs where all his or her doxastic attitudes were supported by his or her total evidence to one where this is not the case, and according to Kopec this would be seen as a decrease in overall epistemic value from the perspective of evidential teleology.

Now, one might question this line of argument. It seems to assume that the positive epistemic value of gaining beliefs that are supported by one's total evidence can easily be

swamped by the negative value of a single belief that is not supported by one's total evidence. We do not see why evidential teleology as stated ought to be committed to that kind of theory of overall value. A different way of respecting the separateness of propositions is, as Berker mentions, to simply build into the theory that doxastic attitudes count as promoting epistemic value only if they promote value for attitudes with the same propositional content. So, on evidential teleology, my doxastic attitude toward p is rational insofar as it promotes my doxastic attitude toward p according with my total evidence.

As we recall, Berker (2013b) points out that veritistic versions of teleology encounter problems with such a modification. Namely, they countenance the kinds of self-fulfilling beliefs on display in the Jane Doe case as rational. However, evidential versions of teleology do not have this problem. Jane Doe's belief that she will recover causes that belief to be true as soon as she adopts it, and therefore has no compensating losses from the perspective of veritism. However, her so believing does not cause the belief to be supported by her total evidence. When she adopts the belief that she will recover against her evidence, her doxastic attitude generates epistemic disvalue and will continue to do so until she actually does recover. Once she actually recovers, it will perhaps be the case that her total evidence now supports a belief to that effect. But given the disvalue generated until that point, the total epistemic value generated by the belief that she will recover will be less than the value generated by the belief that she will not recover, which is at any given time supported by her total evidence.

This concludes our present case for the epistemic rationality of maintaining belief in cases of disagreement when doing so promotes a division of epistemic labor. We have argued that in these cases, strictly traditional epistemic assessment hinders achievement of the goal of inquiry, and faces problems with rational self-defeat and anticipated epistemic regret. Teleological assessment escapes these issues. The verdict that belief in the disagreement case is rational resists the recent arguments against epistemic teleology, as does at least one general teleological norm that can motivate that verdict.

We do not want to overstate the conclusion. While we have argued that belief according to teleological norms better promote the goal of inquiry in these cases, this is only in virtue of promoting our ability to form more rational beliefs as judged by traditional epistemic norms. Furthermore, while we have argued that there is a place for a teleological response to the question of what one ought to believe upon the discovery of disagreement, there is certainly also a prominent place for traditional epistemic assessment in these cases. For example, if we are interested in the impact of disagreement on whether a subject knows that p, or is epistemically

justified in the way that is connected to knowledge, teleological assessment will not help us.

5 The division of epistemic labor and epistemic dependence

We have presented evidence of the benefits of a deliberative division of epistemic labor: deliberation with others with whom we disagree can improve our ability to correctly respond to our evidence. If the above line of argument is correct, it can be epistemically rational to believe in ways that promote these benefits. In this closing section, we consider how dividing our epistemic labor in this way constitutes epistemic dependence on others.

The study of our epistemic dependence is typically employed in cases of testimony. If a speaker knows that p and reliably testifies that p to a hearer, and a hearer competently forms the belief that p as a result, then the hearer can come to know p. Since the hearer does not have other evidence about the truth of p available, he or she epistemically depends on the speaker for his or her knowledge. According to Goldberg (2010), whose theoretical framework is reliabilism, epistemic dependence in such a case goes beyond the more standard ways that features of our environment can matter for the epistemic status of beliefs. In ordinary cases, what determines whether a belief is justified is whether the belief-forming process that caused it is reliable, and the environment is treated as an input to the process. In cases of epistemic dependence, the relevant assessment of reliability must look not only at processes inside the head of the hearer, but also at processes implicated in the speaker's production of the testimony. On Goldberg's view, this is no different from the way that reliability assessments in memorial beliefs must include not just the process of recall that happens as one retrieves a memory trace to form a belief, but also the antecedent processes involved in forming and storing the original belief. In other words, downstream processes inherit the epistemic properties of upstream processes.

While we think it is quite natural to say that two subjects in a case of the deliberative division of epistemic labor also depend on one another epistemically, and that the epistemic assessment of their doxastic attitudes are socially extended, this model will not work in explaining how that is. Consider that in the above case, the dependence relation is entirely asymmetrical. While the epistemic properties of the hearer's belief depend on epistemic properties of the speaker, the epistemic properties of the speaker's belief are independent of any facts about the hearer. But the two agents in a case of the deliberative division of epistemic labor are mutually dependent on one another for the epistemic assessment of their beliefs. In the paradigmatic case we have

discussed, S_1 and S_2 discover that they disagree about p. If they deliberate without first reducing confidence, the ensuing deliberative division of epistemic labor means that S_1 and S_2 land in a better epistemic position. We have argued that this can be rational. The epistemic assessment of S_1's belief in this case is socially extended - the benefits that generate a reason to maintain belief are contingent on S_2 maintaining her belief as well. If either of them reduces confidence while the other maintains confidence, this would not facilitate a deliberative division of epistemic labor, but merely ensure that one side gets a vigorous defense while the other side does not (to the same extent). But it is not the case that what makes S_1's belief rational is that S_2's belief is rational, and that this rationality is transmitted from S_2 to S_1. Neither subject's belief can be epistemically assessed without taking the other's belief into account.

A way to formulate this is to say that unlike the epistemic dependence one finds in cases of testimony, the dependence on finds in the deliberative division of epistemic labor is typically a generative source of epistemic value.[53] Consider Moshman & Geil's (1998) study on the Wason Selection Task. Groups reached consensus on the correct solution after deliberation even when none of the members initially believed that this solution was correct. Epistemic good is generated in the interpersonal process of deliberation, rather than transmitted from one person to another. Humans depend on others epistemically, not only in the sense that we can be beneficiaries of the epistemic labor someone else has done, but also in the sense that we collectively contribute to the generation of epistemic value.

6 Concluding remarks

In this article, we have presented empirical evidence that that motivated reasoning facilitates a deliberative division of epistemic labor in cases disagreement, and that such a division of labor generates epistemic benefits. We have argued that belief can be rational in virtue of promoting these benefits, and pointed toward the need for an expanded notion of epistemic dependence according to which the epistemic evaluations of agents' beliefs is dependent on facts about both agents and their interaction.

[53] Jennifer Lackey and others have argued that testimony and memory *can* function as generative sources of knowledge (Graham, 2006; Lackey, 2005, 2008a). However, in typical cases they rather serve as transmitters of knowledge.

Article 4: Democratic decision-making and the psychology of risk

Coauthored by Andreas Christiansen

1 Introduction

It is a common and immediately plausible thought that, in a liberal-democratic state worthy of the name, the public should play a substantial role in the policy-making process. It is an equally common and plausible thought that, in an enlightened state worthy of the name, policy making should be based on our best understanding of the relevant facts, which in many domains entails that policy making should be based on scientific knowledge. But now a puzzle presents itself: What to do in cases where the public (or large parts of it) want to restrict an activity or technology that they believe to be dangerous, but that scientific experts believe to be safe (or, conversely, where the public is sanguine about an activity or technology that experts believe to be highly risky)? How, if at all, can liberal-democratic and enlightenment values be reconciled? And if they cannot, how should the two conflicting sets of values be balanced?

In order to answer this question well, we need to understand *why* (parts of) the public sometimes disagree with the experts on matters of risk—we need a cognitive and social psychological understanding of public perceptions of risk. And once we have such knowledge, we need to reflect on what implications the psychological facts have for what role the public ought to play in liberal-democratic policy making. These are our two aims in this paper.

In the first part of the paper (section 2), we will present and critically assess the evidence for two major and influential psychological theories of risk perception. One is the bounded rationality theory, according to which (nonexperts') thinking about risk is dominated by the use of fast heuristics that lead to predictable biases in risk perception. The other is the cultural cognition theory, which says that lay beliefs about many risks are a result of culturally (or ideologically) biased processing of evidence, and hence are strongly correlated with cultural (or ideological) worldviews. We will argue that, although both theories have their merits, cultural cognition seems to be at play in a majority of the cases where questions of risk regulation are salient politically.

In the second part of the paper (section 3), we will examine the implications of the psychological theories for three influential liberal-democratic ideas: (A) that public policy should be *responsive* to the preferences of citizens; (B) that liberal-democratic *legitimacy* requires that

policies are reasonably acceptable for all those subject to them; and (C) that the public should directly participate in policy making through public *deliberation*. We will focus on claims made by proponents of each of the psychological theories discussed concerning such implications. In particular, we will engage the views of Cass R. Sunstein, on the side of the bounded rationality theory (Sunstein, 2002a, 2005, 2006), and of Dan M. Kahan, with a number of coauthors, on the side of the cultural cognition theory (Kahan, 2007; Kahan & Slovic, 2006; Kahan et al., 2006).

On Sunstein's view, the fact that public risk perceptions exhibit the biases characteristic of bounded rationality means that they should be disregarded, and that policy should instead be determined by the experts using cost-benefit analysis. We will argue that, although Sunstein is right to point out that bounded rationality undermines the case for being responsive to public preferences for risk regulation, his alternative has its own problems.

According to Kahan and coauthors, the fact that risk perceptions are expressions of cultural or ideological worldviews means that they should be treated much as values are treated in liberal-democratic theory. We will argue that this is largely false. However, cultural cognition theory does contain important insights into how we can overcome the conflict between respecting people's values and respecting the truth when making policy concerning risk.

2. Psychological theories of risk perception

Risk perception research has made it clear that there are a number of domains where a substantial proportion of the public disagree with experts about risk-relevant facts. Genetically modified (GM) foods and global warming are two illustrative examples: according to a report by Pew (Pew Research Center, 2015), 37% of U.S. adults agree that it is safe to eat GM foods, while the corresponding number among AAAS scientists is 88%. 50% of U.S. adults and 87% of AAAS scientists agree that global warming as a result of human activity is occurring, the latter number increasing to 97% among authors of peer-reviewed articles in climate science (Cook et al., 2013, 2016).

The psychology of risk perception aims at explaining such deviations by reference to features of human cognition. The field has been strongly influenced by seminal work by Amos Tversky and Daniel Kahneman on cognitive heuristics and their resulting biases on probability assessments and decision making, as well as their work on prospect theory (Tversky & Kahneman 1974; Tversky & Kahneman 1981; Kahneman 2011). A heuristic is a relatively simple cognitive mechanism that delivers a rapid answer to what may be a complex question, saving time and

cognitive resources. While often accurate, the outputs of heuristics may systematically fail under some circumstances. It is these failures that are denoted as biases. So 'heuristic' refers to a cognitive mechanism while 'bias' expresses a normative assessment of the output of this mechanism, to the effect that something has gone wrong from the point of view of a certain normative theory of reasoning (usually probability theory or logic).

To provide an illustrative example: one of the most well-studied heuristics that is also highly relevant to risk perception is the availability heuristic. When using the availability heuristic to answer a question about the probability of an event, people rely on the ease with which they can recall or imagine instances of such events (Tversky & Kahneman, 1974). While this may usually yield an acceptably accurate estimate, reliance on the availability heuristic leads to systematic biases in the assessment of probability. The probability of highly salient or widely publicized risks, such as tornadoes or homicides, tends to be overestimated, while the probability of less salient risks, such as heart disease or diabetes, tend to be underestimated (Folkes, 1988; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978).

Another heuristic whose more recent discovery had a profound impact on the psychology of risk perception is the affect heuristic (Finucane, Alhakami, Slovic, & Johnson, 2000; Slovic, Finucane, Peters, & MacGregor, 2004). The affect heuristic denotes a tendency for people's judgments of risks and benefits to align along uniformly positive or negative affect towards the risk source. If someone believes that a technology or activity is high risk, she is also likely to believe that its benefits will be low, and vice versa, although there is little reason to suspect that such an inverse correlation usually obtains in reality. This goes beyond people starting with a positive or negative feeling toward a risk source and then generating beliefs about risk and benefits on that emotional background: simply providing people who are naïve with respect to some technology with information that it is high (or low) risk (or benefit) will tend by itself to generate affect, and therefore a belief about benefit (risk) that matches the valence of the initial information. So, if I inform you that a technology, which you currently have no opinion of, is highly risky, this alone will tend to cause you to form the belief that the technology carries little benefit, even in the absence of any direct information about its benefit. More generally, the affect heuristic is representative of an increased awareness within cognitive psychology of the important role emotion plays in risk perception (Sabine Roeser, 2010; Slovic et al., 2004).

Heuristic or emotional information processing is typically cast within a dual process framework where it is contrasted with more deliberate, analytical reasoning (Evans, 2008; Reyna, 2004). When someone is thinking about a technology or activity, a heuristic may yield an initial verdict about risk. Depending on motivation and ability, deliberate reasoning may then be

used to scrutinize and possibly override this initial verdict with one that is the result of more deliberate processing (Evans & Stanovich, 2013). Heuristics that yield strong intuitions or powerful emotional responses are naturally less likely to be overridden.

2.1 Bounded rationality theory

Psychologists are largely in agreement about the above core findings. Nevertheless, there is substantial disagreement about deeper theories of the psychology of risk perception. We first present bounded rationality theory. The term 'bounded rationality' is sometimes used simply to denote that we as humans are subject to limitations in our decision-making apparatus, compared to an ideally rational agent. This is not controversial. What we call *bounded rationality theory* is a more specific series of claims. It holds that our cognitive apparatus aims at providing accurate factual beliefs, but is fallible in achieving this aim because of overreliance on heuristics. When we form a belief about some risk-relevant fact, the function of that belief is to accurately represent some state of affairs to help us make better choices. However, beliefs may fail to fulfil this function because of cognitive limitations. Subjects may lack the time or processing capacity to engage in deliberate reasoning, and therefore rely on heuristics; and since heuristics are vulnerable to biases, our beliefs may be mistaken. These mistakes can be characterized as "blunders" (Sunstein, 2005): they stem from one's acceptance of the output of heuristic processing and failure to engage in sufficient reasoning. When lay people disagree with experts about risk, the reason, according to bounded rationality theory, is that lay people often blunder.[54] They rely on heuristic processing, with their associated biases, in their assessment of risk, whereas experts tend to rely on deliberate reasoning including the scientific method and cost-benefit analysis.

Bounded rationality theory has a wealth of research to support it. It rests largely on the literature on core heuristics such as availability, the affect heuristic, framing, and anchoring—which is extensive and well replicated (Kahneman, 2013; Klein et al., 2014; Shafir & Leboeuf, 2002; Tversky & Kahneman, 1981). Additionally, there is some support to the claim that many mistaken beliefs and bad decisions stem from heuristic processing and that increased deliberate processing tends to predict more accurate beliefs and better decisions. One line of research to provide this support is based on individual differences in rational thought (Stanovich & West,

[54] This is a bit of a simplification. Bounded rationality is also consistent with mistakes being due to a lack of information or to social processes such as information cascades or group polarization (L. R. Anderson & Holt, 1997; Moscovici & Zavalloni, 1969; Sunstein, 2002b).

1998). People who score highly in one type of test of deliberate reasoning tend to score highly in others (Stanovich & West, 2014), and often make better decisions. For example, they tend to make choices under uncertainty that are more utility maximizing compared to people who score low (Frederick, 2005). Another approach is to experimentally impair deliberate reasoning through time pressure or a concurrent cognitive load task, or conversely to force a time delay or otherwise attempt to promote reasoning. Inhibiting reasoning consistently leads to errors and to more impulsive behavior and risk aversion, while bolstering reasoning at least sometimes has the opposite effect (Benjamin et al. 2013).

An aspect of bounded rationality theory that will be important going forward is the implication that people would recognize many of their beliefs as erroneous if they were to engage in the deliberation required to correct their blunder. This hypothetical change of belief might then give rise to different assessments of risk, which would, by virtue of their increased accuracy, be better able to further people's own interests. Thus, adherents of bounded rationality theory can provide a justification for a policy that ignores people's actual beliefs by pointing out that, in addition to better serving their interests, the policy also respects the belief that people actually would have if they were to consider the issue more carefully.

Thus, if the bounded rationality explanation is correct, then we should expect that those parts of the population who disagree with expert judgment about risk-relevant facts do so in part because of a lack of cognitive resources. There are certainly cases where this is borne out. For example, people who tend to rely on intuitive processing profess greater belief in the efficacy of truly ineffectual treatments such as homeopathy to cure disease (Lindeman, 2011). However, questioning the general truth of this prediction is at the heart of the cultural cognition critique of bounded rationality, to which we turn in the next section.

2.2 Cultural cognition theory

As mentioned, there is very little disagreement that humans do rely on heuristics and display biases in their thinking about risk.[55] However, the notion that mistaken factual beliefs as a rule are due to the operation of heuristics has come under strong empirical attack from cultural cognition theory. Cultural cognition theory has its roots in anthropological work that describes

[55] However, the ecological rationality programme of Gerd Gigerenzer and colleagues points out that, far from being a source of ubiquitous bias, heuristics can often be beneficial, providing "fast and frugal" decision procedures that can rival or even beat analytical approaches (Czerlinski, Gigerenzer, & Goldstein, 1999; Gigerenzer & Goldstein, 1996).

societal conflict over risk as structured along two cultural dimensions (Douglas & Wildavsky, 1983). One dimension, individualism-communitarianism, classifies people according to the extent to which they prefer collective solutions to societal problems over individual and market-driven solutions. The other, egalitarianism-hierarchy, describes the extent to which one prefers firmly stratified social orderings in roles and authority. These two dimensions combine into cultural worldviews, which to a large extent predict people's perception of various risk factors depending on their congeniality or lack thereof to the worldview in question. For example, hierarchical individualists will tend to view regulation aimed at industry as questioning the competence of societal elites and the ability of market forces to solve problems, and therefore tend to view the activity of industry as low risk and not requiring such regulation.

This helps explain a feature of risk perception that is hard to make sense of from within a purely bounded-rationality framework: namely, that attitudes toward many risks form coherent clusters that are sharply divided along political and social fault lines. The above-mentioned figure of 50% of U.S. adults affirming the reality of anthropogenic global warming hides a sharp division within the country: the number is only 15% among conservative republicans, but 79% among liberal democrats (Pew Research Center, 2016). Likewise, if one denies the reality of global warming, one is also likely to profess the safety of nuclear power and to favor less gun control. One suggestion from bounded rationality theory might be that this shows one part of the population to be generally more disposed to rely on heuristics than the other. But one would then expect that this group would consistently hold beliefs that are contrary to scientific experts, which is not the case (e.g., as regards the safety of nuclear energy, Pew Research Center, 2015).

To the anthropological base, cultural-cognition theory adds work from psychology on confirmation bias, motivated reasoning, and identity-protective cognition, all of which describe how humans may be biased in their search for, and evaluation of, evidence (Dawson et al., 2002; Kunda, 1990; Nickerson, 1998). Humans tend to seek out and evaluate evidence in ways that are congenial to their believed or desired conclusions. We tend to accept evidence in favor of our favored belief with little scrutiny. If the output of a heuristic bolsters a favored position, then we are unlikely to engage deliberate reasoning to check and possibly overwrite this response. On the other hand, evidence against favored beliefs is heavily scrutinized and subsequently tends to be deemed weak, while heuristic responses that run counter to a favored belief will tend to activate deliberate reasoning in an attempt to find an alternative response (Dawson et al., 2002; Kahan, Peters, et al., 2017; Taber & Lodge, 2006). In evidence-search situations, where people are given the choice between viewing evidence that supports or disconfirms their favored view, subjects tend to select supporting evidence (Jones & Sugden, 2001; Taber & Lodge, 2006).

So, according to cultural cognition theory, cultural worldviews, not costs and benefits, to a large extent determine people's basic attitudes toward various risk sources. These worldviews furnish us with our basic values, which in turn cause us to engage in motivated reasoning in dealing with evidence, with the aim of justifying factual beliefs about these risk sources that protect and bolster the attitude in line with our values.

This suggests a flaw in the bounded-rationality picture. Mechanisms such as motivated reasoning and identity-protective cognition are not heuristics. They are instances of deliberate reasoning, but instances where the aim appears not to be merely a correct appreciation of the facts, but rather to provide support for a particular conclusion. When cultural worldviews are in play during evaluation of evidence regarding a risk source, we are likely to use our reasoning to assess the evidence such that it comes out supporting the position that confirms our worldview. This in turn predicts that widespread increased reliance on reasoning rather than heuristics will not necessarily bring about convergence towards a view closer to the truth. Rather, we should expect those with the greatest propensity and ability to engage in deliberate processing to be *best* at making the evidence yield their favored conclusion (Kahan, 2013).

In an illustrative study (Kahan, Peters, et al., 2017), participants were asked to assess which of two conclusions the results of a (fictional) study supported. In the control version of the task, the study in question was on the efficacy of an experimental crème for the treatment of skin rash. The study's results were presented as a two-by-two matrix, with one dimension denoting whether study subjects' rash got better or worse, and the other denoting whether the subjects had received the treatment or the placebo. Each cell contained a number indicating how many people experienced a certain combination of these dimensions (e.g., people whose rash got better *and* who had received the treatment). Participants had to detect correlation between the variables in order to correctly solve the task. This was so difficult that less than half of participants provided the correct answer (i.e., the result was lower than chance), and performance increased with numeracy (a measure of deliberate processing ability as it applies to numbers and mathematical operations) regardless of cultural background.

In the experimental version of the task, the study was on the effect of gun-control legislation on crime. Here, the cells corresponded to cities that had either implemented a gun-control law recently or not, and whether crime had increased or decreased (e.g., one cell contained the number of cities that had not implemented gun-control *and* had experienced a decrease in crime). Here, a sharp division along cultural lines was seen. If given a version where the correct answer was that crime had decreased as a result of gun control, then liberal participants were likely to find the correct response, and this likelihood increased sharply with

numeracy scores. However, conservative participants given this version were very unlikely to find the correct response, and increased numeracy had no effect on their likelihood to do so. The converse pattern was found for the version where the correct response was that crime had increased: conservatives were quite good at finding the correct response, and highly numerate conservatives much more so than less numerate ones, and liberals were bad at finding the correct response, with increased numeracy offering a very limited benefit. That is, increased capacity to engage in deliberate reasoning helped attaining true beliefs only when the evidence, properly interpreted, was supportive of one's worldview. This suggests that simply providing people with evidence or attempting to engage their deliberative faculty rather than heuristics will do little to correct false beliefs, when these false beliefs are congenial to their cultural worldview. It further suggests that, in general, one should not expect increased deliberative ability to lead to convergence on truth, but rather that one should find the greatest amount of cultural divergence among the most reflective, numerate, and educated.

Research from proponents of cultural cognition theory has borne this out. Across a great many culturally contested domains related to risk, such as global warming, gun control, the HPV vaccine, and fracking, cultural polarization is largest among those with the greatest reflective abilities (Kahan, 2015; Kahan et al., 2012, 2010; Kahan, Peters, et al., 2017). It thus becomes highly problematic to refer to false beliefs that are the result of the mechanisms described by cultural cognition theory as blunders. In many cases, they may be the result of a large amount of deliberate reasoning, rather than an uncorrected heuristic. Likewise, the notion that policy-makers can assume that people's factual beliefs would align with those of scientific experts if only they were to reflect more becomes untenable. What one could expect is rather that increased reliance on deliberate reasoning would lead to belief polarization: more extreme versions of current beliefs (Lord et al., 1979; Taber & Lodge, 2006).

Naturally, far from all domains of risk are culturally contested. For example, there is no cultural conflict over artificial food colorings or sweeteners, cell-phone radiation, the MMR vaccine, or genetically modified foods (in the U.S., although the case may be different in Europe), and in such domains one finds the expected pattern predicted by bounded rationality theory: that higher scientific literacy and reflective capacity increases the likelihood of agreeing with scientific experts, across cultural groups (Kahan, 2015). Thus, one can view cultural cognition theory as describing an important class of exceptions to the general bounded rationality framework rather than as providing a full alternative.

It is an important and, to a large extent, unanswered question for cultural cognition theory why and how certain risks become culturally contested and whether this can be reversed: the

HPV vaccine apparently became culturally salient only following a series of missteps on the part of its manufacturer (Kahan et al., 2010), and even global warming was not a particularly divisive issue in the early 1990s (McCright, Xiao, & Dunlap, 2014).

3 Liberal-democratic decision making

We said at the outset that determining the appropriate balance between relying on experts and including lay citizens' views required understanding what causes citizens to sometimes disagree with experts about what things are risky and what things are safe. We have now seen that the answer is: it's complicated. With respect to some risks, the beliefs of (many) citizens are influenced by heuristics and, as a result, exhibit biases. In those cases, those who are the least scientifically literate and who rely the most on intuitive judgment tend to disagree most with the experts. However, for a substantial number of risks, lay opinion is divided along cultural lines. In these cases, agreement with experts is not correlated with scientific literacy or deliberate, careful reasoning—rather the opposite is true. Instead, an individual's beliefs about the riskiness of some phenomenon largely depends on whether that phenomenon is good or bad according to her basic cultural worldview—her basic values. Furthermore, cases where risk debates have become culturally charged are overrepresented among the risks that exhibit the conflict between experts and (some) citizens, which is our subject in this paper.

So what conclusion can we draw concerning risk management in a state that aims to respect liberal-democratic values and to be enlightened? As noted in the introduction, in assessing the political implications of risk psychology, we will focus on claims that proponents of the two theories we have presented have themselves made. We will structure our discussion according to three core ideas in liberal-democratic political theory. First, there is the idea that public policy should be *responsive* to the preferences of citizens—that is, that differences in public opinion should register as differences in the policies implemented. Second, there is the idea that policies should be such that they could enjoy the assent of all those subject to them. This is most famously engendered in liberal and 'public reason' accounts of political *legitimacy*. And third, there is the idea that the public should directly participate through some form of society-wide *deliberation* on policy issues. We will discuss the implications of the psychological theories for each of these ideas in turn. Before doing so, let us state a couple of clarifications and assumptions.

First, when we are talking about people's risk perceptions in a policy-making context, we are not typically talking about pure factual beliefs. Rather, we are typically talking about one of two things: (i) unprompted exclamations (letters to the editor, demonstrations, etc.) to the effect

that a certain risk is *serious*, an activity is *dangerous*, or that *something must be done* about a risk, or (ii) support, in one form or another, for proposals to regulate the relevant risky activity (e.g., by expressing such support in surveys, by voting for such policies directly in referenda, or by basing one's vote for representative bodies on the risk-regulation platform of the relevant party or candidate). These are (more or less specific) opinions concerning *what policies should be enacted*—they are *policy preferences.*

Second, we will assume that there is in fact consensus among scientific experts concerning a given risk. Note here that *experts'* views of risk are typically not risk perceptions in the sense defined above (i.e., policy preferences). Rather, they are estimates of the probabilities of various (primarily negative) effects of a policy, such as deaths, other health effects, or environmental degradation. We will also assume that (parts of) the public express policy preferences that are at odds with this consensus, in the sense that the following three propositions are true: (a) the public want a technology or another potentially risky thing restricted, (b) this policy preference is based on a belief that the thing in question is risky, and (c) expert consensus is that the thing is not very risky.

3.1 Responsiveness

While it is fairly uncontroversial that it is an ideal of democratic systems that policies are responsive to the preferences of citizens, it is not clear what this ideal entails more precisely. In particular, it is not clear what 'public preferences' means—it might be public opinion as expressed in polls, the preferences expressed by those citizens who actively engage in political debate, or perhaps the preferences policy-makers perceive to be prevalent in the population (See Manza & Cook, 2002, pp. 631–632). Furthermore, it is not obvious what is required for policies to be responsive to such preferences. Typical explications merely hint at an answer, such as that politicians should *take preferences into account* or that policy should be *influenced* by public preferences (Brooks & Manza, 2006, pp. 474–475). How preferences should be taken into account or how much they should influence policy is left open—although most agree that "a perfect correspondence" is neither required nor desirable (Gilens, 2005, p. 778). We want here to set aside debates about what responsiveness is or should be. Instead, we focus on a more basic issue—namely, whether there is even a *prima facie* requirement that the policies of a democratic state should be responsive to citizens' risk perceptions when these are in apparent conflict with expert beliefs.

3.1.1 Sunstein

Sunstein can be seen as arguing that there is no such *prima facie* requirement. At least, he argues that citizens' policy preferences with respect to the regulation of risk-creating activities should play a relatively limited role in policy making. As an alternative, he argues that a major role should be given to cost-benefit analyses performed by experts in regulatory agencies. More precisely, he supports the current (as of 2018) United States system, in which a central agency of the federal government (OIRA, the Office for Informational and Regulatory Affairs) has a mandate to review and reject, on the basis of cost-benefit analyses, regulations suggested by the various technical agencies dealing with environmental, health, and safety policies (such as the Environmental Protection Agency or the Occupational Safety and Health Administration). A main reason for this is a belief that the technical agencies' regulatory priorities reflect public risk perceptions, rather than scientific estimates (Sunstein, 2002a, p. 53, citing Roberts, 1990). The details of Sunstein's proposals are complex, but the main underlying idea is that policy need not be responsive to public risk perceptions, since on his view these are largely (as we have seen above) the products of cognitive biases of various kinds. This conclusion he derives from a general principle: "democratic governments should respond to people's values, not to their blunders" (Sunstein, 2005, p. 126). Since risk perceptions are based on blunders, democratic governments are not required to be responsive to them.

Is he right about this? One possible reason to think that he is not arises if one thinks that the general principle—that democracies should respond only to values, not to blunders—is false. But it is an open question what it would mean for the principle to be false, since it is unclear what the principle says. The problem is that "values" and "blunders" are not exhaustive of the possible descriptions we may give of people's psychological attitudes. True factual beliefs, for example, are clearly neither values nor blunders. Sunstein's principle, then, says that policies *should* be responsive to people's normative beliefs, but *need not* be responsive to their false (or perhaps only obviously false) factual beliefs. This leaves entirely open what we should do when different people or groups hold divergent factual beliefs, none of which is clearly false. In other words, Sunstein's principle has nothing to say about the criteria for selecting which factual beliefs, beyond the clearly false ones, should be allowed to play a role in policy making.

A natural solution to this problem is to add in a principle for selecting respectable factual beliefs. One plausible such principle, congruent with the ideal of enlightened decision making we mentioned in the introduction, would be to use science as a standard-setter. On such a view, any belief conflicting with the scientifically established facts is not entitled to democratic responsiveness. There are ways of questioning this principle, and especially ways of questioning

whether (and how) it could be justified given standard understandings of public reason and the nature of factual disagreements (see, e.g. Jønch-Clausen & Kappel, 2015, 2016). However, we believe the price of giving it up is exceedingly large; since the scientific method is the best known way of generating true factual beliefs, it seems that denying that science can act as gatekeeper for beliefs is tantamount to giving up on having any standards of right and wrong in the empirical domain. So we will accept that beliefs in conflict with established scientific fact are such that democratic governments need not respond to them.

An important caveat needs to be added. In a number of cases, among which are many that are policy relevant, scientific knowledge comes with sizeable uncertainties attached. This needs to be taken seriously by policy-makers. Uncertainty, in effect, means that a number of states of affairs are consistent with the available evidence. In the case of risk, a plausible (but perhaps too simple) way of fleshing this out is to assign only an *interval* of probabilities to a given event, rather than a precise probability (for instance, the probability per year of dying from exposure to pesticides may fall in the interval between one in one million and one in two million). In the case of discrete possibilities—for example, whether gun control works to lower the number of gun-related deaths per year or not—uncertainty means that we cannot believe either discrete possibility very strongly (i.e., the maximum permissible credence for the proposition "gun control works" is relatively close to 0.5). Where uncertainty is involved, the scientific evidence thus does not permit us to give a unique answer to the policy-relevant question—e.g., what the probability per year of dying from pesticide exposure is, or whether gun control works to lower gun-related deaths. Instead, a number of unique answers are possible. It does not fall within the remit of scientific experts to select which of the set of scientifically permissible unique answers to use.

In many cases, however, policy choice depends on what unique answer is correct in the following sense: if p_1 is true, policy R_1 is required (or preferable), but if p_2 is true, R_2 is required. For example, if gun control works, then gun control is (arguably) required—but if gun control does not work, gun control is not required. In such cases, there is a gap between accepting Sunstein's values-not-blunders principle, and delegating decision-making authority to scientific experts, even granting that 'blunders' includes every belief that is contrary to what science says. Public risk perceptions may play some role in filling that gap.

A more important problem with the values-not-blunders principle is that the risk perceptions of ordinary people, being policy preferences, do not straightforwardly fall on either side of the normative-factual belief divide. Consider how an ideally rational person, of the kind one can

meet in decision-theory textbooks, would form her policy preferences concerning a risky activity. Such a person would assign a probability and a value measure ("utility") to each possible outcome of each possible policy, multiply each probability by its utility and sum these products, and advocate the policy that has the highest expected utility. So, even for such a person, a call for a given policy is a consequence of a combination of factual and normative beliefs. Indeed, a policy preference can be made consistent with *any* factual belief, given that the appropriate adjustments are made to the person's normative beliefs. The mere fact that the person calls for a given policy does thus not in itself provide evidence that she has a factual belief that is in conflict with the scientific facts.

However, as we have seen above, the bounded rationality theory that Sunstein relies on provides positive reasons to think that people's factual beliefs concerning risk are often wrong. And (at least to a large extent) the basic fact that nonexperts' beliefs about the magnitude of risks often diverge from the best scientific estimates is not in dispute within psychology. So let us suppose that we can be fairly certain that at least some people have erroneous factual beliefs about the magnitudes of various risks. If it were possible to "implant" true beliefs into such people, then it seems plausible that their risk perceptions (i.e., their more or less precise beliefs about what policies should be enacted) would change.

A very plausible explication of the values-not-blunders principle is then this: what democracy requires is responsiveness to the preferences that people would have had if their factual beliefs were true (or at least not contrary to scientifically established facts).[56] Call this their *counterfactual fact-based preferences*. In so far as policy preferences that ordinary people express currently—call this their *actual preferences*[57]—are different from their counterfactual fact-based preferences, actual preferences are not the kind of thing democracies need to be responsive to.

The normative appeal of this ideal of policy-responsiveness seems to us considerable (although one might want to consider some minimal criteria for what *normative* beliefs are above

---

[56] Discussing the phenomenon of "nudging," where policy proposals have similarly been justified with reference the psychology of heuristics and biases, one observer suggests that the people arguing for such policies "generally believe that social policy should aim to satisfy purified preferences" (Hausman, 2016). "Purified" preferences are preferences people would have had, if they had not been the victims of biases.

[57] Here, and generally in the paper, we use the word 'actually' to indicate what is the case in the *actual* world, as that concept is typically used in possible-worlds ways of speaking of counterfactuals and alethic concepts such as necessity and possibility. That is, we use 'actual' to indicate what is currently the case in the world in which we find ourselves.

board as well). Its main problem is its hypothetical nature. We agree that the ideal form of democratic responsiveness is to the counterfactual fact-based policy preferences of citizens. But in order to implement responsiveness to counterfactual fact-based preferences, we must know (or have reasonably justified beliefs about) what specific preferences a citizen or group of citizens would have had, if they had believed the facts. Note that this is quite a lot harder than having a justified belief that citizens would not have had their actual preferences if they had believed the facts. The real challenge for those who wish to implement responsiveness to counterfactual fact-based preferences is to devise or point to some method for generating reasonably justified beliefs about the specific preferences citizens would have had if they had believed the facts. The only fail-safe way would be to make sure all citizens sincerely believe the facts, to have them determine their policy preferences given those beliefs, and then to make policy responsive to those preferences. But it is of course not possible to run a counterfactual fact-based version of the entire democratic process. So it seems that the best we can aim for is a method that we have reason to believe generates preferences that are reasonable approximations to people's counterfactual preferences.

At least in some places, it seems that Sunstein believes that cost-benefit analysis is a procedure that realizes this. Cost-benefit analysis builds on the approach assumed in decision theory, where (as mentioned above) preferences are a function of separate factual beliefs and value judgments. With respect to factual beliefs, cost-benefit analysis uses the best scientific estimates of the magnitude of risks. As such, it clearly meets the criterion of nonresponsiveness to blunders (although doubts can be had as to whether cost-benefit analysts neglect scientific uncertainty (McGarity, 2002)). With respect to the value judgments, cost-benefit analysts assign a monetary value to a given risk (e.g., a one-in-one-hundred-thousand risk of death per year) based on studies of what people are willing to pay to avoid such a risk, or of what they demand to be paid in order to accept bearing such a risk. Typical ways of measuring willingness-to-pay are studies of wage differentials between risky and safe jobs, and surveys asking people directly for their valuations. Sunstein suggests that "the governing theory" behind this approach "follows [people's] own judgments about risk protection" (Sunstein, 2014, p. 86). Although he also stresses that the current practice does not fully realize the governing theory—in particular, it does not sufficiently take into account differences in risk valuations across individuals—he seems to believe that the general willingness-to-pay approach measures people's own valuations of a given risk (as he says, "the limitations [of current theory] are practical ones" (Sunstein, 2014, p. 136)). By combining these valuations with the facts and assuming the framework of

decision theory, cost-benefit analysis arrives at the preferences people would have had if they had believed the facts.

The idea that the methods of cost-benefit analysis tracks people's own valuations—their counterfactual fact-based preferences—is not universally accepted. It relies on extrapolation of behaviors in one context, in particular the labor market, to all other contexts, and on assumptions from economics and rational choice theory that are in many ways questionable (see, e.g., E. Anderson, 1993, ch. 9; Hausman, McPherson, & Satz, 2017, ch. 9). Furthermore, the very same biases and heuristics that Sunstein is eager to expel from risk management through the use of scientific estimates are likely to influence people's valuations of risks in willingness-to-pay studies. Finally, survey studies frequently register a large number of so-called protest valuations (where people state a willingness to pay either nothing or an implausibly large amount, or perhaps decline to state a number at all), indicating a rejection of the very idea of using willingness to pay as a valuation measure for public goods (Kahneman, Ritov, Jacowitz, & Grant, 1993). Such responses are typically disregarded, which suggests that cost-benefit analysis is ill equipped to deal with preferences that are not of the type typically relevant in markets. Thus it does not succeed in capturing the counterfactual fact-based preferences of those who reject treating a given policy domain as appropriately governed by the ideals of a market economy.

The conclusions that can be drawn from the above are limited. We have merely suggested that Sunstein's proposal of delegating much of the policy-making power to scientific experts doing cost-benefit analyses is not plausibly an ideal solution to risk regulation. So even if Sunstein is right that risk perceptions—of the unfiltered kind that are expressed in the various more or less precise calls for risk-regulating policies—are too tainted by their partial source in cognitive biases to be taken into account in policy making, his alternative may not be much better. At least, his alternative does not embody ideal responsiveness (i.e., responsiveness to counterfactual fact-based preferences). It is doubtful that ideal responsiveness *can* be fully realized in practice. It may be the case that the available realizable alternatives leave us with a dilemma: if we make policy responsive to expressed risk perceptions, we will be *overresponsive* to false or unscientific beliefs; but if we make policy unresponsive to these risk perceptions, we will be *underresponsive* to values. In other words, the seemingly simple ideal of responsiveness to values and nonresponsiveness to blunders may be an unattainable ideal. Call this the *responsiveness dilemma.*

3.1.2 Cultural cognition

Kahan and his coauthors argue that cultural cognition theory further undermines Sunstein's approach. Recall first what the cultural cognition theory says about how people form risk perceptions. On the cultural cognition model, risk perceptions are not formed in the way assumed by decision theorists (and by Sunstein)—that is, by combining pure factual beliefs about the numerical magnitude of risks (expected deaths, probabilities of ecosystem damage, and the like) with pure normative beliefs about how bad the various possible bad effects of a policy are. Instead, people assess (probably mostly unconsciously) the relationship between a possibly risky activity and their cultural worldview—and thus assess at the same time whether *restricting* the activity is justifiable, or perhaps required, according to their view of the ideal society. Thus, as we mentioned above, hierarchical individualists balk at regulation of industry because it questions the competence of elites (hierarchy) and assumes the inadequacy of market solutions (individualism). Conversely, egalitarians dislike the activity of capitalist industry generally, and thus welcome restrictions. Based on such general assessments of the value of activities and of restrictions on them, people form factual as well as normative beliefs about the risks and benefits of the activity, in a kind of post-rationalization procedure, in which motivated assessment of evidence concerning the effects of the activity and policy is central.[58] Consequently, "citizens invariably conclude that activities that affirm their preferred way of life are both beneficial and safe, and those that denigrate it are both worthless and dangerous," and even the factual aspect of risk perceptions (could they be isolated) "express [citizens'] worldviews" (Kahan et al. 2006, p. 1105).

Kahan et al. argue that cultural cognition theory undermines Sunstein's view in two related ways. First, they claim that Sunstein's strategy of using cost-benefit analysis to realize the values-not-blunders ideal "borders on incoherence" (Kahan et al., 2006, p. 1105). In other words,

[58] There are two likely mechanisms at play: First, people form beliefs about whether a given type of risk regulation is desirable, based directly on their cultural worldview. That is, there is a direct causal link from worldviews to policy preferences. Second, people form *factual* beliefs—about the numerical magnitude of risks—through motivated cognition, wherein cultural worldviews affect people's assessment of the evidence concerning the riskiness (or safety) of an activity. Here, the causal link goes from worldviews to assessment of evidence, and thus to pure factual beliefs, and then in a second step from those factual beliefs to policy preferences. Since pure factual beliefs are not easily disentangled from policy preferences (see, e.g., Kahan & Slovic, 2006, pp. 166-168), it is difficult to test which of these mechanisms is the dominant one. However, in a study of self-defense cases, Kahan and Braman found support for the view that "the influence that values exert over outcome judgments is mediated by the impact of the commitments on individuals' perceptions of the facts" (Kahan & Braman, 2008, p. 45)—i.e., for the second mechanism.

the fact that risk perceptions are due to cultural cognition means that the cost-benefit approach does not realize the ideal embodied in the values-not-blunders principle. On one reading, this would merely be the claim we have just made: that cost-benefit analysis fails to respect values. But of course this would be completely independent of the cultural cognition theory. The values we have argued are overridden in cost-benefit analysis are ordinary normative beliefs (about the value of a human life, say), not culturally influenced factual beliefs (about how many lives a certain activity will claim). Second, they suggest that "bringing the role of cultural cognition into view severely undermines the foundation for Sunstein's refusal to afford normative significance to public risk evaluations generally" (Kahan et al., 2006, p. 1004). That is, they suggest that acknowledging the role of cultural cognition undermines the case for nonresponsiveness to citizens' actual policy preferences.

How might the fact that people's risk perceptions are shaped by cultural cognition further undermine the cost-benefit analysts' approach and/or strengthen the case for responsiveness to actual preferences? We suggest that cultural cognition points to two different facts that may be important: (1) that the relationships between values (in the form of cultural worldviews), factual beliefs, and policy preferences are not as Sunstein and others assume, and (2) that risk perceptions are rooted in cultural worldviews, and therefore are expressions of citizens' values.

Let us first consider issue (1). Here, the claim would be that the fact that risk perceptions are due to cultural cognition means that they do not behave in ways that Sunstein and others assume—for example, that changes in factual beliefs do not change preferences in the way assumed—and that this undermines the strategy of cost-benefit analysis further and/or strengthens the case for responsiveness to actual preferences. Such a claim could be made in two ways:

(i) Since both factual beliefs and policy preferences are due to the same underlying cause, we should not expect changes in factual beliefs to change policy preferences. As Kahan et al. put it,

> "risk perceptions originating in cultural evaluation are not ones individuals are likely to disown once their errors are revealed to them. Even if individuals could be made to see that their cultural commitments had biased their review of factual information … they would largely view those same commitments as justifying their policy preferences regardless of the facts. (Kahan et al. 2006, p. 1105)"

On this reading, an individual's counterfactual fact-based preferences are likely to be the same as his actual preferences (i.e., the preference he would hold if he believed the facts is likely to be the same as the preference he currently holds). If that is the case, people's actual preferences are at least a good approximation of their counterfactual fact-based preferences. Thus we have a solution to the problem of how to achieve responsiveness to counterfactual fact-based preferences—namely, to use actual preferences. Or, to put the matter differently, it is not true that responsiveness to actual preferences is overresponsiveness to faulty factual beliefs, since actual preferences are not influenced by factual beliefs at all—faulty or not. Reading (i) would, then, give reason to be responsive to citizens' actual preferences.

Reading (i) faces two problems. The first problem is that the claim that changes in factual beliefs do not change policy preferences seems too strong, and it goes beyond what can be justified by the evidence that the cultural cognition theory relies on. Cultural cognition is primarily a thesis about how cultural commitments lead to biased assessment of evidence, such that one believes the evidence supports the factual beliefs that fits one's cultural commitments best. But it is possible to debias people at least to some degree, and to bring them towards mutual agreement on the facts. And furthermore, there is evidence that such debiasing alters people's policy preferences, bringing previously opposed parties closer together (Cohen et al., 2007). So it seems to us that the fact of cultural cognition does not justify ignoring the problem of overresponsiveness to false beliefs.

The second problem is that, at least in many policy domains, preferences may lose some of their claim to democratic responsiveness if they turn out to be too resistant to the facts. Resistance to changes in factual beliefs may reveal policy preferences to be based in kinds of value judgments that are unacceptable from a liberal-democratic point of view—e.g., a desire to regulate purely private behavior (such as sexual behavior or harmless commercial activities) or worldviews that deny the fundamental equality of all citizens (such as racist or sexist views). If it were the case that citizens' policy preferences would not change regardless of what the facts are, we would at least need to examine the substantive content of those preferences in more detail—and to reserve judgment as to whether those preferences merit democratic responsiveness until we have a better understanding of what that substantive content is.

(ii) Since policy preferences and factual beliefs are both caused by people's cultural worldviews (i.e., their most basic values), any change in factual beliefs requires a change in basic values. Suppose a given citizen actually has faulty factual beliefs, and that these beliefs are due to cultural cognition. According to reading (ii), the basic values this citizen actually holds are *not*

the basic values she would hold in the counterfactual case where she came to believe the facts. The cost-benefit analysts' method is essentially an attempt to disentangle actual factual beliefs from actual value judgments. The analysis then recombines actual value judgments with the *true* facts, and thereby generates a policy preference. But on reading (ii), such an approach does not succeed in revealing citizens' counterfactual fact-based preferences. The cost-benefit method uses a citizen's *actual* values, but cultural cognition shows that these are likely to be different from her *counterfactual fact-based* values. In other words, a citizen's counterfactual fact-based preferences are not (as Sunstein believes) a function of her actual values and the facts, but a function of a *new* set of values and the facts.

Reading (ii) would show that the cost-benefit analysts' method does not successfully track people's counterfactual fact-based preferences. It also suggests that it is difficult to predict how people's preferences would change if they sincerely came to believe facts that are in conflict with their cultural worldviews. Thus it lends support to the use of more deliberative methods, wherein real flesh-and-blood people are allowed to undergo a change in their views in response to facts and arguments (unlike methods like cost-benefit analysis, which seeks to *infer* what people would prefer from data about what they actually believe, value, and prefer). Consequently, the "deliberative debiasing" methods Kahan et al. argue in favour of using are supported by this reading (Kahan et al., 2006, pp. 1100-1104).

Kahan et al.'s other claim—that cultural cognition supports responsiveness to actual preferences—is not supported by reading (ii). At best, reading (ii) shows cost-benefit analysis to be a worse approximation of the ideal of responsiveness to counterfactual fact-based preferences than we might otherwise have thought. However, this merely makes the responsiveness dilemma worse, by making one of the horns of that dilemma worse. It is not obvious, however, that reading (ii) is of much help in deciding how to choose when faced with a responsiveness dilemma—that is, if we have to choose between responsiveness to actually expressed preferences and (something like) cost-benefit analysis.

Let us now move to issue (2), the fact that cultural cognition theory shows risk perceptions to be expressions of values. Kahan et al. state that "when expert regulators reject as irrational public assessments of the risks associated with putatively dangerous activities … they are in fact overriding *values*" (Kahan et al. 2006, p. 1105). It is, unfortunately, not clear what is meant by "public assessments of … risks" in this quote. On the one hand, the phrase might refer to policy preferences, such as that a given activity A is dangerous and should be regulated. On the other

hand, it might refer to people's purely factual beliefs about the magnitude of risks. Let us now consider each of these two readings of issue (2) in turn (we call them readings (iii) and (iv) to avoid confusion with (i) and (ii) above):

(iii) *Experts are overriding G_1's values because they implement a policy R_2 that is different from G_1's preferred policy R_1.* Recall that the kind of case we are interested in has the following structure: (a) the public wants a technology or another potentially risky thing restricted, (b) this policy preference is based on a belief that the thing in question is risky, and (c) expert consensus is that the thing is not very risky. In the group-based framework of cultural cognition, 'the public' should be replaced with some cultural group. So we assume that a cultural group G_1 wants the activity A restricted through policy R_1, and that G_1 wants this because they believe *p*, that A carries certain risks. The experts, based on sound science, believe ¬*p* (i.e., that A does not carry those risks) and therefore implement a policy R_2 that does not restrict A appreciably.

In cases of this kind, it is hard to see why we should accept that implementing a policy other than R_1 overrides G_1's values. By assumption, G_1 prefers R_2 *because* they believe *p*—the implication being that they would not have preferred R_1 if they had believed ¬*p* (i.e., that R_1 is not their counterfactual fact-based policy preference). Once more, there are now two possibilities for what G_1's policy preference would then have been if they had believed ¬*p*. First, G_1 might have preferred, or at least acquiesced to, R_2, the policy implemented by the experts. In that case, the expert decision procedure would have achieved its ideal aim. Thus there would be no reason to be responsive to G_1's actual preference, and we would have no reason to object to the experts' decision procedure either. Second, G_1 might have preferred some third possible policy R_3. In that case, we would still have no reason to demand that policy be responsive to G_1's actual preferences. However, there would be reason to complain that the experts' decision procedure has failed to be responsive to G_1's values. Insofar as we cannot tell *a priori* whether G_1 would have preferred (or acquiesced to) R_2 or not, the conclusion that follows is that we cannot be confident that the experts' decision procedure is responsive to G_1's values, in the absence of some effort to determine what G_1's counterfactual fact-based preferences are.

But perhaps the assumption that G_1 prefers R_1 *because* they believe *p* is not correct. That is, perhaps the case is one in which G_1 would prefer R_1 regardless of the facts—G_1's factual belief that A is dangerous is merely a *post hoc* rationalization of the group's policy preference, which it holds for other reasons than that A is dangerous. Kahan and Braman (2008, pp. 51-54) suggest that it is only in cases of this kind—where people would not alter their policy preference

even if they came to believe the facts—that there is a demand for policy responsiveness to preferences. At the same time, however, they speculate that people would *not* be inclined to hold on to their preferences if they were to realize that their factual beliefs are the product of cultural cognition, at least in the case they are discussing (cases of self-defense). The same might well be the case in typical instances of risk regulation. In the case where people *would* hold on to their policy preferences after coming to believe the facts, the problem we mentioned under reading (i) above recurs: G_1's preference for R_1 has some basis other than that A is in fact risky, and that basis may show the preference to be less reasonable than it initially seemed.

Consider, for example, the case of regulation of industry pollution. Recall that hierarchical individualists tend to be skeptical of such regulation because it casts doubt on the competence of societal elites and the ability of market forces to solve problems, and consequently tend to believe that the risks associated with industry pollution are low. But suppose hierarchical individualists were brought to sincerely believe that some industry's emission of a certain chemical C creates severe risks to the health of those exposed, but that they persisted in their policy preference (not to regulate). What could the basis of such that preference then be, other than a blatant disregard for the welfare of those who will likely suffer health problems? A similar problem arises for egalitarians, who are inclined to approve of restrictions of "commerce and industry, which they see as sources of unjust social disparities" (Kahan, 2012, p. 728), and who consequently tend to believe that the risks associated with industry pollution are high. Suppose egalitarians persisted in their desire to regulate emissions of C even after having sincerely accepted that C does not pose a serious risk to anyone. The only possible basis of such a preference is then a general anti-industry agenda. By persisting in their preferences, both the hierarchical individualists and the egalitarians would violate basic norms of risk regulation, such as that people have some right to be protected against serious risks and that harmless private behavior cannot be restricted.

Thus it seems to us that in the case of risk regulation there is reason to be skeptical of policy preferences that would not change if people were to come to believe the facts. So, while the possibility that policy preferences would not change if people came to believe the facts does provide some reason to be responsive to those preferences, there will simultaneously be a reason not to be responsive. However, in cases where people merely *overestimate* risk (or underestimate, as the case may be), persisting in policy preference is less problematic. It may reflect, for example, a judgment that the aim of protecting people's health is very important relative to the aim of securing favorable conditions for business. But this is just the general

problem with cost-benefit analysis we identified above. It is not obvious that the phenomenon of cultural cognition adds much to that problem.

(iv) Experts are overriding G_1's values by denying the pure factual beliefs of G_1 (i.e., p*), since those factual beliefs express values.* Since believing *p* is an expression of G_1's values, the validity of G_1's values is denied when expert regulators implement a policy based on the fact that ¬*p* is true. We think the view that merely denying (a group of) citizens' factual views is to be underresponsive to their values has both strange and dangerous implications. Suppose, for example, that the experts in this case implement G_1's preferred policy R_1, but also believe (and state publicly) ¬*p*. On the view considered, the implication would be that the experts' policy making is insufficiently responsive to the values of G_1 in this case, even though G_1 got its preferred policy implemented. That seems to us a strange implication, which requires an excessive demand for responsiveness.

Alternatively, consider a case like the one we mentioned above, where G_1 would at least acquiesce to the expert's implementation of R_2 if they were to come to believe the truth (i.e., ¬*p*). One might think that, since the belief *p* is an expression of G_1's values, implementing R_2 exhibits a lack of responsiveness to G_1's values even though R_2 is G_1's counterfactual fact-based preference (or a least would be acceptable to G_1 in those counterfactual circumstances). In effect, this would amount to denying that policy preferences that unequivocally depend on factual beliefs that do not meet the required correctness criterion (i.e., beliefs that are blunders or contrary to scientifically established facts) do not merit democratic responsiveness. This seems to us a dangerous implication. In factual matters, priority must be given to the truth, and to our best methods for finding out the truth. And in fact, Kahan et al. seem to share our worry here. In a response to Sunstein's response to their original paper, Kahan and Slovic "admit to a fair measure of ambivalence about when beliefs formed as a result of cultural cognition merit normative respect within a democratic society," and concede that "if we came off sounding as if we think democracy entails respecting all culturally grounded risk perceptions, no matter how empirically misguided they might be, we overstated our position" (Kahan & Slovic, 2006, pp. 170–171).

In conclusion, Kahan et al.'s skepticism towards Sunstein's proposed use of expert cost-benefit analysis is largely warranted, but it is questionable if the fact of cultural cognition contributes much to the problems with cost-benefit analysis. To be sure, cultural cognition provides a

different set of reasons for thinking that cost-benefit analysis does not succeed in tracking counterfactual fact-based preferences—but arguably that claim was already very well supported by other reasons. Furthermore, cultural cognition theory provides only very limited reason to be responsive to actual preferences in cases where these are in conflict with experts' scientific assessments of the riskiness of an activity. Cultural cognition theory therefore does not warrant solving the responsiveness dilemma in favor of responsiveness to actual preferences. It does, however, provide support for using deliberative debiasing techniques to solve that dilemma.

3.2 Liberal legitimacy

We now move from the democratic to the liberal aspect of the liberal-democratic ideal—more precisely, to the liberal conception of legitimacy. According to this conception, political power is legitimate only if could be reasonably accepted by all subject to it. While many philosophers are attracted to some version of the liberal legitimacy principle, there is no general agreement on what the principle precisely amounts to. It is controversial how demanding the requirement that political power be acceptable to all is—does it require that all can accept the basic procedure by which laws and policies are made (Rawls's (1993) view) or does it require that each law or policy be reasonably acceptable to all? The latter is obviously a much more demanding criterion. It is likewise controversial how demanding the reasonability clause is—should our conception of reasonability be such that the acceptance of most people as they really exist is required, or do we need to secure acceptance only from people whose views meet higher standards of justifiability? And there are more conflicts as well (for an overview, see Quong, 2018).

Kahan et al. suggest that the cultural cognition theory does have important implications for how policy may be made if it is to be legitimate on the liberal conception. On Kahan et al.'s explication of the liberal ideal, it consists in an "injunction that the law steer clear of endorsing a moral or cultural orthodoxy" (Kahan et al, 2006, p. 1106). They then go on to suggest that "it is questionable whether risk regulation should be responsive to public demands for regulation, since these express cultural worldviews"—that is, exactly the kind of views that it would be wrong for policy to endorse according to the liberal ideal. So even though Kahan et al. seem to believe that the dubious factual basis of risk-related policy preferences is not sufficient to strip them of their claim to democratic responsiveness, they suggest that there are *liberal* reasons for making policy nonresponsive to such preferences.

Kahan et al. do not elaborate what they mean by "endors[ing] a moral or cultural orthodoxy." But since they cite the writings of Bruce Ackerman and John Rawls in support of

the principle, let us assume that the following, common liberal idea is what Kahan et al. have in mind: legitimacy requires policies to be justified only with reference to reasons that are public, in the sense that all reasonable citizens agree that these reasons count in favor of (or against, as the case may be) policies. Now suppose we have identified an exhaustive set of such reasons, and that these are the only ones actually given weight in the policy-making process. Obviously policies at the same time will reflect factual assumptions about how much various policies realize the values defined by public reasons. If the cultural cognition theory is correct, factual assumptions are not value neutral, since each set of factual assumptions expresses a cultural worldview.

What is the import of this for liberal legitimacy? The basic question is what it means that factual assumptions express worldviews and when that would be a problem. Suppose a policy is justified only on the basis of public reasons and the facts. In that case, it seems to us strange to say that the policy in question is illiberal merely because the facts are (coincidentally) endorsed by adherents of one cultural worldview. 'Expressing a worldview' must refer to something more substantial than this kind of *correspondence to* a worldview if it is to be a liberal problem. This reflects the basic assumption we endorsed earlier—namely, that the facts, and scientific methods of establishing facts, ought to have priority in policy making.

Perhaps the problem arises only in cases where there is genuine uncertainty about what the facts are. Suppose that the scientific evidence concerning gun control allows for believing either that gun control does prevent deaths from firearm accidents and crimes (call this p) or that gun control does not prevent such deaths ($\neg p$).[59] And suppose further that the public reasons bearing on the case are such that if p is true, then gun control should be implemented, and if $\neg p$ is true, gun control should not be implemented (e.g., because there is a presumption of liberty). So policy *must* endorse either p or $\neg p$, in the sense that one policy follows from p and a different policy follows from $\neg p$. Supposing that p reflects the cultural worldview of one group G_1 and that $\neg p$ reflects the worldview of G_2, it seems that policy must endorse one group's worldview although the other group's view is not in conflict with science.

Suppose that one thinks that basing policy on either of p or $\neg p$ would be illiberal. Such a view would run into the following problem: it is a plausible requirement for any criterion of legitimacy that at least one policy is legitimate. But in the example given here, we must either

[59] Kahan's own treatment of this case (2007, pp. 120-122) seems to imply that this is the case. However, more recent evidence suggests that gun control does, in fact, lower gun-related injuries and fatalities (Santaella-Tenorio, Cerdá, Villaveces, & Galea, 2016).

say that *both* policies are legitimate or that *neither* policy is legitimate, since they are symmetrically situated with respect to their basis in both public reasons and factual assumptions. Since the view that neither policy is legitimate is not a viable option, we must say that both policies are legitimate. Consequently, G_1 does not have a viable complaint that a no-gun-control policy is illegitimate, although it does in one sense express the cultural worldview of G_2—and similarly G_2 has no legitimacy complaint against gun control.

Another possible interpretation of what it means that a policy preference expresses a worldview is that the worldview is the *real, causal explanation* for why a certain person or group has the preference. On this reading, calls for regulation of a given risk, although seemingly justified by reference to public reasons, are really caused by "an unjust desire to use the expressive capital of the law for culturally imperialist ends" (Kahan et al., 2006, p. 1107). Suppose the policy in question is above board in the sense that some combination of public reasons and scientifically acceptable factual assumptions would justify the policy. Would the fact that this legitimate rationale is not the real reason why the policy is implemented constitute a legitimacy problem? The assumption here is that the group implementing the policy sincerely (and correctly) believes that the policy has a legitimate rationale, a fact that they exploit in order to implement a policy that they desire in any case. Such a group could be accused of an unattractive opportunism. But this does not constitute a legitimacy problem on the orthodox interpretations of the liberal legitimacy criterion.[60] The liberal criterion stresses the importance of all groups *being able to* reasonably accept the policy. Since the policy here is *ex hypothesi* justifiable based on a set of normative assumptions and a set of factual assumptions, both of which are reasonable (i.e., the set of public reasons and the set of scientifically accepted facts, respectively), all groups are able to reasonably accept the policy. It would be unreasonable for a group to demand that the factual assumptions best expressing *their* worldview be the basis of the law rather than another set of reasonable factual assumptions.

We conclude, then, that the fact that factual beliefs express cultural worldviews in the way the cultural cognition theory has revealed does not entail any problems from the point of view of the

[60] There is some debate among theorists of public reason regarding the appropriate role of *sincerity*. Some views within this debate hold it to be a requirement for legitimacy that public reasons are the actual motivation for people's advocacy of a given policy (see Schwartzman, 2011, pp. 387–390). Kahan et al. may of course defend their position by endorsing such a view, but in doing so they would no longer be able to claim the support of the liberal principle of legitimacy *tout court*.

liberal conception of legitimacy in cases where policies are justifiable based on reasonable normative and factual beliefs.

3.3 Deliberation

In the previous section we discussed public reason as (a part of) a substantive account of policies' legitimacy. We were thus concerned with whether a certain class of reasons provide sufficient justification for a policy. But 'public reason' is also frequently used to refer to a certain norm of deliberation. Here, the concern is not so much whether a policy *could* be justified with reference to agreed-upon, public reasons, but what reasons we may make appeal to in the process of policy making—in public and parliamentary debate, in the civil service, and in courts. According to the deliberative norm of public reason, citizens, politicians, judges, and others may appeal only to reasons that are neutral between reasonable conceptions of the good. The idea, then, is to remove all appeals to contested worldviews from the public arena.

Kahan (2007) takes issue with this public-reason norm. On Kahan's reading, the public-reason norm has two rationales: First, it *disciplines* those in power by demanding that they pursue only policies that they sincerely believe are supported by public reasons. And second, it *protects* those out of power by ensuring that laws are such that they can accept them without thereby denouncing their vision of the good life (Kahan, 2007, p. 129). But, according to Kahan, the cultural cognition theory reveals that the public-reason norm fails to produce either of its promised effects. The demand for secular justifications does not prevent those in power from imposing their vision of the good on society, since even the sincerely held belief that a policy promotes the public good reflects a cultural worldview. And the demand does not ensure that political losers accept policies enacted by their opponents either. More likely, they will interpret opponents' arguments for those policies as disingenuous and reflecting a "smug insistence of their adversaries that such policies reflect a neutral and objective commitment to the good of all citizens" (Kahan, 2007, p. 131).

Kahan suggests that the public-reason norm be replaced with its polar opposite, which he calls the "expressive overdetermination" norm. According to this norm, justifications of policies in the public forum should not avoid references to contested worldviews and conceptions of the good—they should instead attempt to show how the relevant policy promotes the substantial cultural commitments of all groups. Casting this in Rawlsian terms, we might say that the desire for overlapping consensus among adherents of rival comprehensive views should not lead us to

ban reference to the content of these comprehensive views—say, to religious values, strongly egalitarian ideals, or free-market principles. Instead we should attempt to show that all of these values, in all their comprehensive thickness, support some policy (Kahan, 2007, pp. 131-132). The proposal builds on research from social psychology on self-affirmation. The kinds of biases in processing of evidence highlighted by cultural cognition theory stem from a motivation to defend one's identity by defending factual beliefs perceived to be important to the groups with which one identifies. Self-affirmation research has shown that these defensive motivations, and therefore the biases, are decreased when aspects of subjects' personal or social identities are affirmed—for example, by allowing them to write a brief essay outlining a value or group membership that is important to them. In effect, affirmation provides an identity 'buffer' such that one can afford to lower one's cognitive defenses. People whose identities have been affirmed are thus more objective in assessing evidence and arguments, either written or during discussions (Cohen et al., 2007; Cohen, Aronson, & Steele, 2000; Correll, Spencer, & Zanna, 2004; Sherman & Cohen, 2002). Expressive overdetermination takes advantage of this: highlighting that a policy is in line with the values of one's group is taken to be one way of affirming the importance of that value. If so, one can expect people to subsequently be less biased in assessing the risks and benefits of the policy. Thus, expressive overdetermination is meant to achieve the goals of having public policy recognized by all groups as legitimate, and of diffusing the intensely conflictual nature of politics.

(Kahan et al., 2015) provide direct evidence that expressive overdetermination may be effective. Hierarchical individualists were more likely to rate a study concluding that extant emission limits would be insufficient to avoid environmental catastrophe as valid and to express that climate change posed a high risk if they had previously been exposed to a study suggesting that geoengineering was a necessary element in combating climate change. Since geoengineering does not involve imposing restrictions on free enterprise or suggest that corporate elites are unable to solve collective problems, this framing highlighted that the reality of climate change need not threaten hierarchical individualist values. In fact, these values were affirmed insofar as a privately driven use of technology was cast as necessary to combat climate change. This allowed hierarchical individualists to assess the evidence more objectively without threat to their identity.

The realization that seemingly conflict-diffusing mechanisms, such as the public-reason norm, may in fact not work—or may even be counterproductive—seems to us to be the most directly

useful insight for political philosophy that follows from the understanding of cultural cognition. Nevertheless, we do have some misgivings about the expressive-overdetermination norm and about Kahan's dismissal of the public-reason norm.

Let us start with the latter. Is it really true that the public-reason norm fails to deliver on both of its promises? First, consider whether the norm disciplines those in power. The cultural cognition theory shows that the mere fact that those in power *sincerely believe* policies to be supported by public reasons does not ensure that policies *are in fact* so supported. However, it remains plausible that the public-reason norm *contributes* to the aim of liberally legitimate policies. The mere demand that evidence that a certain policy promotes publically recognized goods must be produced will likely provide some constraints on what policies will be implemented by conscientious adherents to the public-reason norm. Although processing of evidence is culturally biased as described above, there are limits to the degree to which people can pick the evidence that suits them (Kunda, 1990). Furthermore, there is evidence (Cohen et al., 2007; Luskin et al., 2012; Vinokur & Burnstein, 1978) that deliberations between adherents of conflicting worldviews or ideologies brings these people closer together with respect to their factual beliefs. Insofar as the willingness (and perhaps even active desire) to engage with the arguments of political opponents is also a part of the public-reason norm, it has resources to diffuse the kind of conflicts that arise from cultural cognition as well.

Second, consider the protective aim of public reason. A corollary of the above is that the public-reason norm does not plausibly increase the likelihood that liberally illegitimate policies will be enacted (rather, it plausibly lowers that likelihood). So there is no reason to think that losers are less well protected under the public-reason norm than in the case where appeals to "thick" values can freely be made. What the cultural cognition theory shows with respect to losers is that they are likely to feel aggrieved even when they have no right to do so (since policies are legitimate). So only if the goal is to ensure *actual* acceptance on the part of losers does the public-reason norm fail. This is a worthy goal, but less important than protecting them from illiberal cultural imperialism.

Now what about the expressive-overdetermination norm as an alternative? Supposing that Kahan accepts the standard public-reason account of *legitimacy*, expressive overdetermination does not contribute to the legitimacy of policies. On that account, a policy that is *in fact* justifiable by reference to public reasons only is legitimate. The fact that a group falsely believes that a policy is not so justifiable does not alter the fact that it is. Furthermore, expressive overdetermination does not contain any resources that increase the likelihood of policies that are in fact legitimate, or any resources that lower the likelihood of policies that are not legitimate.

There are nonstandard accounts of public reason that may be more conducive to seeing expressive overdetermination as having a legitimacy-creating role. On the convergence view of Gerald Gaus, for example, legitimacy requires that each citizen be able to support the policy from within her own total view (Gaus, 2011; Gaus & Vallier, 2009). Gaus's main argument for viewing legitimacy in this way is that reasons that people hold as part of their comprehensive view, but which are not public reasons, may *defeat* the justification of a policy based on public reasons. Consequently, people would not be able to sincerely accept the imposition of that policy. This line of argument meshes well with the protection function of deliberative norms as Kahan describes it. However, the convergence view faces the potential problem that there will often not be a policy that can gain support from all comprehensive points of view. Additional principles for determining what policies are legitimate in such cases are then needed. Gaus has developed an elaborate theory for this purpose, but nothing Kahan has written suggests that he would go along with Gaus in this regard. If a legitimacy-incurring role for expressive overdetermination is to be grounded in an account like Gaus's, much work remains to be done to flesh out the theory.

Return now to more standard accounts of public reason. Since expressive overdetermination does not contribute to policies' legitimacy, it seems that the expressive-overdetermination norm can be justified only instrumentally, as a means to an end. The most immediately obvious end that the norm serves is to ensure actual acceptance of policies by all groups. And actual acceptance is presumably valuable because it realizes the aims of disciplining the powerful and protecting the powerless. But there is some reason to be skeptical that actual acceptance will realize those goals. Expressive overdetermination can be used to secure acceptance from groups without substantially respecting their values. Consider an example that Kahan points to—namely, French abortion law. This law gives women access to abortions, but in order to secure acceptance from conservatives, this access is available only in an "emergency" (Kahan, 2007, p. 132). However, no criteria for what constitutes an emergency were included, and no questioning of a woman's own declaration that an emergency exists is allowed. In effect, then, the emergency clause is substantively empty, and was included only for its symbolic meaning. While this construction did succeed in creating a consensus on the policy, it is hard to see why those who believe in any serious way that non-emergency abortions is a problem *should*

have been satisfied with this law.[61]

On the other hand, expressive overdetermination might be used for another end—namely, to enable people holding conflicting views to converge on the facts (cf. the climate change study described above), and hence to diffuse or avoid cultural conflict over factual questions. Kahan et al. have provided strong evidence that the public-reason norm does not realize this goal particularly well, and that a norm of expressive overdetermination can (perhaps somewhat counterintuitively) realize the goal better. However, and as Kahan himself recognizes,[62] expressive overdetermination is merely one tool for achieving fact convergence.

4 Conclusion

We have argued above that the psychological facts of risk perception are complex. Divergences between experts and lay citizens are sometimes at least partly a reflection of lack of scientific literacy and overreliance on heuristics on the part of some citizens. But in other cases, cultural worldviews seem to be behind differences of opinion over what is risky and what is not. And in fact those seem to be the cases that are most interesting politically, such as global warming, environmental issues, or GM foods (in Europe).

However, we have also argued that the fact that faulty beliefs express people's basic values has few implications for how liberal-democratic states should go about formulating policy with respect to putatively risky activities and technologies. Contrary to what proponents of cultural cognition argue, the fact that risk perceptions express cultural worldviews does not give us stronger reasons than we would otherwise have for making policy responsive to such perceptions. Similarly, the fact that factual beliefs about risks express visions of the ideal society does not undermine the legitimacy of using scientifically accepted facts as the basis for policy making.

This largely means that we are stuck with the responsiveness dilemma that we identified in our discussion of Sunstein's view: if policy is insulated from the people, we risk being underresponsive to citizens' values, and if policy is made in a more populist manner, we risk overresponsiveness to false beliefs. However, the cultural cognition theory does provide some important insights into how this dilemma can be resolved. It supports the case for using

[61] Of course, one might not think that the anti-abortion party's views were such that they ought to be respected on a liberal view of legitimacy—but the example is illustrative of the risk that expressive overdetermination can be used to manipulate groups to accept policies that illegitimately trample their values.

[62] http://www.culturalcognition.net/blog/2014/4/5/cognitive-illiberalism-expressive-overdetermination-a-fragme.html (comment by Dan Kahan, aka. "dmk38").

structured deliberation methods to determine what citizens' preferences would be if they were to come to accept scientific facts. And it provides significant guidance for those of us who want to reform political discourse in a way that enables reasonable discussion of policies based on common acceptance of the relevant facts.

Conclusions and perspectives

The conclusions of the articles in this thesis, particularly those three pertaining to the epistemology of disagreement, might seem to point in disparate directions. Article 1 concluded that motivated reasoning, and the belief polarization that is sometimes its outcome, are epistemically irrational when the agent is aware that the proposition is controversial. On the other hand, article 3 concluded that this very same type of reasoning about controversial propositions facilitates a division of epistemic labor that is conducive to the rationality of deliberating groups with internal disagreement, and that this can make the individual rational in maintaining what may very well be a belief that has a history of such polarization. Similarly, article 2 concluded that political disagreement has epistemic significance, and that our evidence in these cases supports a reduction of confidence (although there is a puzzle about how to determine its magnitude). But article 3 seems to suggest that we should not reduce confidence in these cases if we anticipate collective deliberation.

Initial appearances to the contrary, the three conclusions are not in any strict conflict with one another. Articles 1 and 2 base their conclusions on the broadly evidentialist framework that figures in standard discussions in the epistemology of disagreement. They are about what your evidence supports in cases of disagreement, and the permissible ways of processing this evidence. Article 3 relies for its conclusion on a departure from this framework. It was argued that even if your evidence supports a reduction of confidence, all things considered epistemic rationality might not require doxastic revision in response to such evidence when it carries significant epistemic costs.

What the three articles, taken together, suggest is that social context is key to determining the epistemic rationality of our doxastic attitudes. The same type of reasoning and the same doxastic attitude can be irrational when we are alone, but rational when we reason together. We might then consider motivated reasoning an instance of Mandevillian intelligence: A case where a mode of reasoning that, taken in isolation, is individually irrational, is conducive to collective rationality (Smart, 2017). This result should perhaps not be all too surprising. Psychologists have proposed that the primary evolutionary function of reasoning, that is, the reason why such a metabolically expensive and, as we have seen, not always epistemically reliable, mechanism evolved, is to facilitate the evolution of communication (Mercier & Sperber, 2011).

Communication can have enormous benefits for both speakers and hearers, but its evolution requires that hearers are able to filter out false claims and attempts at deception (Krebs & Dawkins, 1984). The ability to check whether a conclusion follows from, or is made more likely by, a series of premises serves this function when hearers are not willing to take a conclusion on trust. The presence of skeptical listeners in turn generates a selective pressure for speakers to present persuasive arguments for their conclusions if they want to be acknowledged. This has resulted in a reasoning mechanism that is inclined to motivated reasoning and confirmation bias – when your aim is to persuade someone, your had best generate reasons in favor of your conclusion rather than against it. But it has also resulted in collective reasoning that appears remarkably well adapted to arriving at correct interpretations of evidence.

A pertinent question is to what extent the teleological notion of rationality can impact not just cases of disagreement immediately followed by collective deliberation, but also cases where any deliberation occurs further into the future. If I engage in motivated reasoning in isolation, this might result in my generating reasons that could conceivably benefit group deliberation down the road, and that I would not have generated (to the same extent) had I been guided by accuracy goals. If it were to turn out that my being motivated in my reasoning now ends up promoting my adopting a doxastic attitude that is supported by the evidence via collective deliberation in the future, to what extent does this impact the rationality of my initial motivated reasoning and its resulting belief? It seems plausible that some manner of temporal discounting of future epistemic benefit is in order when determining epistemic rationality on teleological grounds, but it is not at all clear how sharp this discounting function should be.

In light of the epistemic benefits of deliberation with those we disagree with, we might also be interested in whether there could be *epistemic* reason for us to act is such ways as to promote the occurrence of such deliberation (Booth, 2006). The results suggest that if we want to arrive at beliefs that are supported by our evidence, we should actively seek out those with whom we disagree and engage them in discussion. While the notion of epistemic reason for action departs quite far away from the kinds of concerns that epistemologists typically focus on, I think it is worth exploring carefully the possibility that we might have such reasons.

Now for a worry: The empirical work cited in article 3 to document the epistemic benefits of deliberation in groups with internal disagreement largely focuses on disagreements that are not political in nature or tied to our personal or social identities in a strong way. There is an unfortunate dearth of studies that apply similar experimental setups to investigate collective

reasoning about these types of issues. This means that there is a risk that the kinds of disagreement addressed in articles 1, 2, and 4, would not benefit from collective deliberation to the same extent. Perhaps when it comes to politics, we are unwilling to change our minds in response even to the strongest reasons presented to us in collective deliberation, and this would preclude any consensus on the doxastic attitude best supported by the evidence. I think there are grounds for modest optimism that the results would generalize. We do typically change our doxastic attitudes in response to strong arguments even when it comes to issues we feel strongly about or that tie into our identities, particularly when we reason in dialogue (Mercier, 2011; Park, Levine, Westerman, Orfgen, & Foregger, 2007; Paxton, Ungar, & Greene, 2011). And group deliberation specifically has been shown to change minds: When we deliberate with those we disagree with, whether it be face to face or on social media, we typically move closer to their view, and gain a better understanding of their reasons, even when the topic of discussion is highly value-laden or politically charged (Barberá, 2015; Cohen et al., 2007; Luskin et al., 2012; V. Price, Cappella, & Nir, 2002; Vinokur & Burnstein, 1978), although there are some exceptions to this pattern in the literature (Wojcieszak & Price, 2010). Of course, doxastic change is not necessarily indicative of epistemic benefit, so there remains a critical need for studies that directly address whether groups are better able to arrive at what is actually supported by the evidence in cases of political or otherwise charged disagreements.

A somewhat related worry about the generalizability of the results of article 3 pertains to whether the normative argument transfers to the political domain. Recall that in article 1, it was shown that motivated reasoning does not always depend on prior belief. Sometimes what looks like motivated reasoning in the defense of a prior belief can actually be motivated reasoning in defense of a value or commitment with which the factual belief is entangled. Even in the absence of prior belief, motivated reasoning can shape the subsequent processing of information. But this might be taken to suggest that, when it comes to political beliefs, there are actually no costs to conciliating, because motivated reasoning will proceed regardless of whether one abandons the prior belief and so facilitate collective deliberation. It is hard to experimentally manipulate beliefs and directly observe what effect (or lack thereof) conciliating would have on subsequent reasoning, but it is certainly a possibility that motivated reasoning could remain even after conciliating. This would, I think, be sufficient to show that there actually is no conflict between conciliation and epistemic benefits in political cases. I think this would be a happy outcome, as I am inclined to be sympathetic to conciliatory views of what our evidence supports in cases of disagreement, and so I could have my cake and eat it too. But I nevertheless maintain that the

studies on the downstream effects on reasoning of political and ethical 'belief reversal' from the choice blindness experiments mentioned in article 3 suggest that there may be such a conflict.

The results of article 3 also suggest the need for what we might call a 'social epistemic game theory'. Space did not allow for a full exploration of this issue in the article itself, but the final section noted that, on the view put forth, whether I am rational in maintaining confidence in the face of disagreement depends on whether you do the same. Aikin et al. (2010) have pointed out the unsavory consequences for a conciliatory agent when other agents do not play by the same rules. Our result would seem to be vulnerable to a similar kind of problem, although on our view, both agents would maximize their epistemic rationality by not reducing confidence. Cases like this nevertheless suggest the need for a fuller exploration of the how the rationality of the beliefs of various agents are conditional not only on their own doxastic 'actions' but also those of other agents, and what kinds of broader socio-epistemic norms we might seek to promote in order to promote 'epistemic cooperation'.

Another question in need of further elucidation is what features of groups with internal disagreement and their individual members are conducive to good deliberation. Woolley et al. (2010, 2015) offer a promising step in this direction in studies showing that, in addition to diversity, group performance is correlated with the proportion of women in the group (an effect that is driven by emotional perceptiveness, which women tend to be superior with respect to), and with the degree of equality of turn-taking in speaking. In contrast, the mean IQ of group members is, surprisingly, not a good predictor of group performance. Along similar lines, Aikin & Clanton (2010) suggest that we should develop what they call group-deliberative virtues: individual intellectual virtues that improve our ability to contribute to fruitful group deliberation, such as deliberative wit (as opposed to dullness), deliberative friendliness (as opposed to quarrelsomeness), or deliberative humility (as opposed to arrogance). An interesting feature of (some of) these virtues is that their desirability is going to be specific to the context of deliberation as well. The notion of deliberative wit, for example, likely contains many of the psychological mechanisms that cause us to individually polarize when we consider evidence, whereas it is beneficial in a group context.

More broadly speaking, the results referred to in articles 2 and 4 about the relationship between cognitive ability and polarization raise the question of what kinds of intellectual qualities we should seek to impart in citizens. We saw that constructs that are typically considered

unequivocally good, such as science literacy, reflectiveness, and open-mindedness, contribute to the polarization of beliefs about facts that are relevant to politics. While this may not go so far as to suggest that we should not want citizens to be scientifically literate, it does suggest that it will not be sufficient that they are if we want them to converge on the view supported by our best scientific evidence. Other virtues need to be promoted in tandem. A very promising candidate to mention here is curiosity, which, to the best of my knowledge, is alone among the intellectual virtues that have been studied in quantitative political psychology in predicting *decreases* in polarization, and convergence in the direction supported by expert opinion and the scientific evidence (Kahan, Landrum, Carpenter, Helft, & Jamieson, 2017). And, if the results from article 3 turn out to hold up in the political domain, another candidate would be a virtue that consists in an inclination to actively seek out disagreement and engage in collaborative discussion with our political opposites.

References

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Psychology*, *29*(2), 97–119.

Ahlstrom-Vij, K., & Dunn, J. (2014). A Defence of Epistemic Consequentialism. *Philosophical Quarterly*, *64*(257), 541–551.

Aikin, S. F., & Clanton, C. J. (2010). Developing Group-deliberative Virtues. *Journal of Applied Philosophy*, *27*(4), 409–424. http://doi.org/10.1111/j.1468-5930.2010.00494.x

Aikin, S. F., Harbour, M., Neufeld, J. A., & Talisse, R. B. (2010). Epistemic Abstainers, Epistemic Martyrs, and Epistemic Converts. *Logos and Episteme*, *1*(2), 211–219.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*(6), 1037–1049. http://doi.org/10.1037/h0077720

Anderson, E. (1993). *Value in ethics and economics*. Cambridge, MA.: Harvard University Press.

Anderson, L. R., & Holt, C. A. (1997). Information Cascades in the Laboratory. *American Economic Review*, *87*(5), 847–862. http://doi.org/10.2307/2951328

Andreoni, J., & Mylovanov, T. (2012). Diverging opinions. *American Economic Journal: Microeconomics*, *4*(1), 209–232. http://doi.org/10.1257/mic.4.1.209

Andreou, C. (2017). Dynamic choice. In *Stanford Encyclopedia of Philosophy* (Spring 201). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/dynamic-choice/

Asch, S. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70.

Bagg, S. (2015). Can deliberation neutralise power? *European Journal of Political Theory*, *0*(0), 1–23. http://doi.org/10.1177/1474885115610542

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, *91*(4), 612–625. http://doi.org/10.1037/0022-3514.91.4.612

Baliga, S., Hanany, E., & Klibanoff, P. (2013). Polarization and Ambiguity. *The American Economic Review*, *103*(7), 3071–3083.

Ballantyne, N. (2015). Debunking Biased Thinkers (Including Ourselves). *Journal of the*

American Philosophical Association, *1*(1), 141–162. http://doi.org/10.1017/apa.2014.17

Barberá, P. (2015). How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S. *APSA Conference Paper*.

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology*, *32*(1), 176–184. http://doi.org/10.1037/h0076771

Baumeister, R. F., Ainsworth, S. E., & Vohs, K. D. (2016). Are Groups More or Less than the Sum of their Members? The Moderating Role of Individual Identification. *Behavioral and Brain Sciences*, *39*, e137.

Benjamin, D., Brown, S. a, & Shapiro, J. (2013). Who is "Behavioural"? Cognitive Ability and Anomalous Preferences. *Journal of the European Economic Association*, *11*(6), 1231–1255.

Berker, S. (2013a). Epistemic Teleology and the Separateness of Propositions. *Philosophical Review*, *122*(3), 337–393. http://doi.org/10.1215/00318108-2087645

Berker, S. (2013b). The rejection of epistemic consequentialism. *Philosophical Issues*, *23*(Epistemic Agency), 363–387.

BonJour, L., & Sosa, E. (2003). *Epistemic justification: internalism vs. externalism, foundations vs. virtues*. Blackwell Pub.

Booth, A. R. (2006). Can there be epistemic reasons for action? *Grazer Philosophische Studien*, *73*(1), 133–144.

Boudry, M., & Braeckman, J. (2012). How Convenient! The Epistemic Rationale of Self-validating Belief Systems. *Phillosophical Psychology*, *25*(3), 1–21.

Bratman, M. E. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press.

Bratman, M. E. (2012). Time, rationality, and self-governance. *Philosophical Issues*, *22*(1), 73–88. http://doi.org/10.1111/j.1533-6077.2012.00219.x

Bratman, M. E. (2014). Temptation and the Agent's Standpoint. *Inquiry*, *57*(3), 293–310. http://doi.org/10.1080/0020174X.2014.894271

Brogaard, B. (2014). Wide-scope requirements and the ethics of belief. In J. D. Matheson & R. Vitz (Eds.), *The Ethics of Belief*. Oxford University Press.

Brooks, C., & Manza, J. (2006). Social Policy Responsiveness in Developed Democracies. *American Sociological Review*, *71*(3), 474–494. http://doi.org/10.1177/000312240607100306

Carr, J. R. (2017). Epistemic Utility Theory and the Aim of Belief. *Philosophy and*

Phenomenological Research, *95*(3), 511–534. http://doi.org/10.1111/phpr.12436

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199596195.001.0001

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *Philosophical Review*, *116*(2), 187–217. http://doi.org/10.1215/00318108-2006-035

Christensen, D. (2009). Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass*, *4*(5), 756–767. http://doi.org/10.1111/j.1747-9991.2009.00237.x

Christensen, D. (2010). Higher-Order Evidence. *Philosophy and Phenomenological Research*, *LXXXI*(1), 185–215. http://doi.org/10.1111/j.1933-1592.2010.00366.x

Christensen, D. (2014a). Conciliation, Uniqueness and Rational Toxicity. *Noûs*, *0*(3), n/a-n/a. http://doi.org/10.1111/nous.12077

Christensen, D. (2014b). Disagreement and Public Controversy. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 1–33). Oxford University Press.

Christensen, D., & Lackey, J. (2013). *The epistemology of disagreement: new essays*. (D. Christensen & J. Lackey, Eds.). Oxford University Press.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. D. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology*, *93*(3), 415–430. http://doi.org/10.1037/0022-3514.93.3.415

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self. *Personality and Social Psychology Bulletin*, *26*(9), 1151–1164. http://doi.org/10.1177/01461672002611011

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. Oxford University Press.

Cook, J., & Lewandowsky, S. (2016). Rational Irrationality: Modeling Climate Change Belief Polarization Using Bayesian Networks. *Topics in Cognitive Science*, *8*(1), 160–179. http://doi.org/10.1111/tops.12186

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., … Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt K B, Tignor M, M. H. L. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2). http://doi.org/10.1088/1748-9326/8/2/024024

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., … Rice, K. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4).

http://doi.org/10.1088/1748-9326/11/4/048002

Corner, A., Whitmarsh, L., & Xenias, D. (2012). Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation. *Climatic Change*, *114*(3–4), 463–478. http://doi.org/10.1007/s10584-012-0424-6

Correll, J., Spencer, S. J., & Zanna, M. P. (2004). An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology*, *40*(3), 350–356. http://doi.org/10.1016/j.jesp.2003.07.001

Cowie, C. (2014). In defence of instrumentalism about epistemic normativity. *Synthese*, *191*(16), 4003–4017. http://doi.org/10.1007/s11229-014-0510-6

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? *Simple Heuristics That Make Us Smart*. http://doi.org/10.1002/(SICI)1099-0771(200004/06)13:2<161::AID-BDM348>3.0.CO;2-P

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated Reasoning and Performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387. http://doi.org/10.1177/014616702236869

DeScioli, P., Massenkoff, M., Shaw, A., Petersen, M. B., & Kurzban, R. (2014). Equity or Equality? Moral Judgments Follow the Money. *Proceedings of the Royal Society B*, *281*(1797), 20142122.

Douglas, M., & Wildavsky, A. B. (1983). *Risk and culture : an essay on the selection of technological and environmental dangers*. University of California Press.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *The Behavioral and Brain Sciences*, *38*, e130. http://doi.org/10.1017/S0140525X14000430

Dunn, J. (2013). Peer Disagreement and Group Inquiry. *Annual Meeting of the Indiana Philosophical Association.*

Easwaran, K., Fenton-glynn, L., Hitchcock, C., & Velasco, J. D. (2016). Updating on the Credences of Others. *Philosophers' Imprint*, *16*, 1–39.

Elga, A. (2005). On overrating oneself... and knowing it. *Philosophical Studies*, *123*(1–2), 115–124. http://doi.org/10.1007/s11098-004-5222-1

Elga, A. (2007). Reflection and Disagreement. *Noûs*, *41*(3), 478–502. http://doi.org/10.1111/j.1468-0068.2007.00656.x

Enoch, D. (2010). Not Just a Truthometer: Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement. *Mind*, *119*(476), 953–997. http://doi.org/10.1093/mind/fzq070

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm.

Psychological Bulletin, *128*(6), 978–996. http://doi.org/10.1037//0033-2909.128.6.978

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. http://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory Cognition*, *11*(3), 295–306.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. http://doi.org/10.1177/1745691612460685

Feldman, R. (2006). Epistemological puzzles about disagreement. In S. Hetherington (Ed.), *Epistemology Futures* (pp. 126–236). Oxford: Oxford University Press.

Feldman, R., & Warfield, T. A. (2010). *Disagreement*. Oxford University Press. Retrieved from http://www.amazon.co.uk/Disagreement-Richard-Feldman/dp/0199226083

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17. http://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S

Firth, R. (1981). Epistemic Merit, Intrinsic and Instrumental. *Proceedings and Addresses of the American Philosophical Association*, *55*(1), 5–23.

Foley, R. (2001). *Intellectual Trust in Oneself and Others*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511498923

Folkes, V. S. (1988). The Availability Heuristic and Perceived Risk. *Journal of Consumer Research*, *15*(June). http://doi.org/10.1007/978-94-007-1433-5

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. Retrieved from http://www.jstor.org/stable/4134953

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*(April), 1–12. http://doi.org/10.1016/j.jesp.2017.04.003

Fumerton, R. (2010). You Can't Trust a Philosopher. In R. Feldman & T. A. Warfield (Eds.), *Disagreement* (pp. 91–110). Oxford University Press. http://doi.org/10.1093/acprof

Funk, C. L., Smith, K. B., Alford, J. R., Hibbing, M. V., Eaton, N. R., Krueger, R. F., … Hibbing, J. R. (2013). Genetic and Environmental Transmission of Political Orientations. *Political Psychology*, *34*(6), 805–819. http://doi.org/10.1111/j.1467-9221.2012.00915.x

Gaus, G. F. (1996). *Justificatory liberalism: an essay on epistemology and political theory*. Oxford University Press.

Gaus, G. F. (2011). *The order of public reason: a theory of freedom and morality in a diverse and bounded world*. Cambridge University Press.

Gaus, G. F., & Vallier, K. (2009). The roles of religious conviction in a publicly justified polity. *Philosophy & Social Criticism*, *35*(1–2), 51–76. http://doi.org/10.1177/0191453708098754

Gelfert, A. (2011). Who is an Epistemic Peer? *Logos and Episteme*, *2*(4), 507–514.

Gerber, A., & Green, D. (1999). Misperceptions About Perceptual Bias. *Annual Review of Political Science*, *2*(1), 189–210. http://doi.org/10.1146/annurev.polisci.2.1.189

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650–669. http://doi.org/10.1037/0033-295X.103.4.650

Gilens, M. (2005). Inequality and Democratic Responsiveness. *Public Opinion Quarterly*, *69*(5), 778–796. http://doi.org/10.1093

Goldberg, S. C. (2010). *Relying on Others: An Essay in Epistemology*. Oxford University Press. http://doi.org/10.1093/acprof

Goldberg, S. C. (2011). The Division of Epistemic Labour. *Episteme*, *8*, 112–125. http://doi.org/10.3366/epi.2011.0010

Goldberg, S. C. (2017). Should have known. *Synthese*, *194*(8), 2863–2894. http://doi.org/10.1007/s11229-015-0662-z

Goldman, A. I. (2015). Reliabilism, Veritism, and Epistemic Consequentialism. *Episteme*, *12*(2), 131–143. http://doi.org/10.1017/epi.2015.25

Graham, P. J. (2006). Can Testimony Generate Knowledge? *Philosophica*, *78*, 105–127.

Greaves, H. (2013). Epistemic decision theory. *Mind*, *122*(488), 915–952. http://doi.org/10.1093/mind/fzt090

Greitemeyer, T., Schulz-Hardt, S., Brodbeck, F. C., & Frey, D. (2006). Information sampling and group decision making: the effects of an advocacy decision procedure and task experience. *Journal of Experimental Psychology: Applied*, *12*(1), 31–42. http://doi.org/10.1037/1076-898X.12.1.31

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE*, *7*(9), e45457. http://doi.org/10.1371/journal.pone.0045457

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. *PLoS ONE*, *8*(4), e60554. http://doi.org/10.1371/journal.pone.0060554

Hallen, B. L., Bingham, C. B., Hill, C., Carolina, N., & Cohen, S. L. (2017). At Least Bias is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives. *Available at SSRN: https://ssrn.com/abstract=2952510*, *April 3*. http://doi.org/10.1007/s10551-015-2769-z.For

Hamilton, L. C. (2011). Education, politics and opinions about climate change: evidence for interaction effects. *Climatic Change*, *104*(2), 231–242. http://doi.org/10.1007/s10584-010-9957-8

Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy*, *82*(7), 335–349.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling Validated Versus Being Correct: A Meta-Analysis of Selective Exposure to Information. *Psychological Bulletin*, *135*(4), 555–588. http://doi.org/10.1037/a0015701

Hausman, D. M. (2016). On the Econ within. *Journal of Economic Methodology*, *23*(1), 26–32. http://doi.org/10.1080/1350178X.2015.1070525

Hausman, D. M., McPherson, M. S., & Satz, D. (2017). *Economic analysis, moral philosophy, and public policy* (3rd ed.). Cambridge: Cambridge University Press.

Hawthorne, J., & Srinivasan, A. (2013). Disagreement without transparency: Some bleak thoughts. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement: New Essays* (pp. 9–30). Oxford: Oxford University Press.

Heath, Y., & Gifford, R. (2006). Free-Market Ideology and Environmental Degradation: The Case of Belief in Global Climate Change. *Environment and Behavior*, *38*(1), 48–71. http://doi.org/10.1177/0013916505277998

Hedden, B. (2015). Time-Slice Rationality. *Mind*, *124*(494), 449–491. http://doi.org/10.1093/mind/fzu181

Hennes, E. P., Ruisch, B. C., Feygina, I., Monteiro, C. A., & Jost, J. T. (2016). Motivated Recall in the Service of the Economic System : The Case of Anthropogenic Climate Change. *Journal of Experimental Psychology: General*, *145*(6), 755–771. http://doi.org/10.1037/xge0000148

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *The Behavioral and Brain Sciences*, *37*(3), 297–307. http://doi.org/10.1017/S0140525X13001192

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from chilhood to adolescence:*

an essay on the construction of formal operational structures. Routledge.

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. http://doi.org/10.1177/1745691612464056

Isenberg, D. J. (1986). Group Polarization : A Critical Review and Meta-Analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why Differences Make a Difference: A Field Study of Diversity, Conflict, and Performance in Workgroups. *Administrative Science Quarterly*, *44*(4), 741. http://doi.org/10.2307/2667054

Jenkins, C. S. (2007). Entitlement and rationality. *Synthese*, *157*(1), 25–45. http://doi.org/10.1007/s11229-006-0012-2

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–24. http://doi.org/10.1037/a0035941

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–9. http://doi.org/10.1126/science.1111709

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99. http://doi.org/10.1023/A:1005296023424

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375. http://doi.org/10.1037/0033-2909.129.3.339

Jost, J. T., Nosek, B. A., & Gosling, S. D. (2008). Ideology. Its Resurgence in Social, Personality, and Political Psychology. *Perspectives on Psychological Science*, *3*(2), 126–136.

Jønch-Clausen, K., & Kappel, K. (2015). Social Epistemic Liberalism and the Problem of Deep Epistemic Disagreements. *Ethical Theory and Moral Practice*, *18*, 371–384. http://doi.org/10.1007/s10677-014-9523-y

Jønch-Clausen, K., & Kappel, K. (2016). Scientific Facts and Methods in Public Reason. *Res Publica*, *22*(2), 117–133. http://doi.org/10.1007/s11158-015-9290-1

Kahan, D. M. (2007). The Cognitively Illiberal State. *Stanford Law Review*, *60*, 115. http://doi.org/10.2307/40040378

Kahan, D. M. (2012). Cultural Cognition as a Conception of the Cultural Theory of Risk. In S. Roeser, R. Hillerbrand, M. Sandin, & M. Peterson (Eds.), *Handbook of Risk Theory* (pp. 725–759). Springer. http://doi.org/10.1007/978-94-007-1433-5

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and*

Decision Making, *8*(4), 407–424.

Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology*, *36*(S1), 1–43. http://doi.org/10.1111/pops.12244

Kahan, D. M. (2016). The Politically Motivated Reasoning Paradigm Part 1: What Politically Motivated Reasoning Is and How to Measure It. *Emerging Trends in Social & Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, 1–16.

Kahan, D. M. (2017). The expressive rationality of inaccurate perceptions. *Behavioral and Brain Sciences*, *40*, e6. http://doi.org/10.1017/S0140525X15002332

Kahan, D. M., & Braman, D. (2008). The Self-Defensive Cognition of Self-Defense. *American Criminal Law Review*, *45*(1), 1–65.

Kahan, D. M., Braman, D., Cohen, G. L., Gastil, J., & Slovic, P. (2010). Who fears the HPV vaccine, who doesn't, and why? an experimental study of the mechanisms of cultural cognition. *Law and Human Behavior*, *34*(6), 501–16. http://doi.org/10.1007/s10979-009-9201-0

Kahan, D. M., Braman, D., Gastil, J., Slovic, P., & Mertz, C. K. (2007). Culture and identity-protective cognition: Explaining the white male effect in risk perception. *Journal of Empirical Law Studies*, *4*(3), 465–505. http://doi.org/10.1111/j.1740-1461.2007.00097.x

Kahan, D. M., Braman, D., Slovic, P., Gastil, J., & Cohen, G. L. (2009). Cultural cognition of the risks and benefits of nanotechnology. *Nature Nanotechnology*, *4*(2), 87–90. http://doi.org/10.1038/nnano.2008.341

Kahan, D. M., & Corbin, J. C. (2016). A note on the perverse effects of actively open-minded thinking on climate-change polarization. *Research & Politics*, *3*(4). http://doi.org/10.1177/2053168016676705

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174. http://doi.org/10.1080/13669877.2010.511246

Kahan, D. M., Landrum, A. R., Carpenter, K., Helft, L., & Jamieson, K. H. (2017). Science Curiosity and Political Information Processing. *Advances in Political Psychology*, *38*(Suppl. 1). http://doi.org/10.1111/pops.12396

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54–86. http://doi.org/10.1017/bpp.2016.2

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate

change risks. *Nature Climate Change*, *2*(10), 732–735. http://doi.org/10.1038/nclimate1547

Kahan, D. M., Silva, C. L., Tarantola, T., Jenkins-Smith, H., & Braman, D. (2015). Geoengineering and Climate Change Polarization: Testing a Two-channel Model of Science Communication. *Annals of the American Academy of Political & Social Science*, *658*(92), 192–222. http://doi.org/10.2139/ssrn.1981907

Kahan, D. M., & Slovic, P. (2006). Cultural Evaluations of Risk: "Values" or "Blunders"? *Harvard Law Review*, *119*, 166–172. http://doi.org/10.1525/sp.2007.54.1.23.

Kahan, D. M., Slovic, P., Braman, D., & Gastil, J. (2006). Fear of Democracy: a cultural evaluation of Sunstein on risk. *Harvard Law Review*, *119*(4), 1071–1109.

Kahan, D. M., & Stanovich, K. E. (2016). Rationality and Belief in Human Evolution. *Annenberg Public Policy Center Working Paper No. 5*. Retrieved from https://ssrn.com/abstract=2838668

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *The American Psychologist*, *58*(9), 697–720. http://doi.org/10.1037/0003-066X.58.9.697

Kahneman, D. (2013). *Thinking, Fast and Slow*. Turtleback Books.

Kahneman, D., Ritov, I., Jacowitz, K. E., & Grant, P. (1993). Stated Willingness to Pay for Public Goods: A Psychological Perspective. *Psychological Science*, *4*(5), 310–315. http://doi.org/10.2307/40063053

Kappel, K. (2017). Bottom Up Justification, Asymmetric Epistemic Push, and the Fragility of Higher Order Justification. *Episteme*, 1–20. http://doi.org/10.1017/epi.2017.19

Kelly, T. (2002). The Rationality of Belief and Some Other Propositional Attitudes. *Philosophical Studies*, *110*(2), 163–196. http://doi.org/10.1023/A:1020212716425

Kelly, T. (2005). The Epistemic Significance of Disagreement. In J. Hawthorne & T. S. Gendler (Eds.), *Oxford Studies in Epistemology. Volume 1* (pp. 1–36). Oxford: Oxford University Press.

Kelly, T. (2008). Disagreement, Dogmatism, and Belief Polarization. *Journal of Philosophy*, *105*(10), 611–633.

Kelly, T. (2010). Peer disagreement and higher order evidence. In R. Feldman & T. A. Warfield (Eds.), *Disagreement* (pp. 183–217). Oxford University Press.

Kelly, T. (2013). Disagreement and the Burdens of Judgment. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.

Kelly, T. (2016). Evidence. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Winter 201).

Kelly, T., & White, R. (2014). Evidence Can(not) be Permissive. *Contemporary Debates in*

Epistemology, 312–323.

Kelp, C. (2014). Two for the knowledge goal of inquiry. *American Philosophical Quarterly*, *51*(3), 227–232.

Kenyon, T. (2014). False polarization: debiasing as applied social epistemology. *Synthese*, 2529–2547. http://doi.org/10.1007/s11229-014-0438-x

King, N. L. (2012). Disagreement: What's the Problem? Or A Good Peer is Hard to Find. *Philosophy and Phenomenological Research*, *85*(2), 249–272. http://doi.org/10.1111/j.1933-1592.2010.00441.x

Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy*, *87*(1), 5–22.

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*(4), 852–884. http://doi.org/10.1037//0033-295X.107.4.852

Klausen, S. H. (2009). Two Notions of Epistemic Normativity. *Theoria*, *75*(3), 161–178. http://doi.org/10.1111/j.1755-2567.2009.01045.x

Klein, R. a., Ratliff, K. a., Vianello, M., Adams ., R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. a. (2014). Investigating Variation in Replicability. *Social Psychology*, *45*(3), 142–152. http://doi.org/10.1027/1864-9335/a000178

Kopec, M. (2012). We Ought To Agree: a Consequence of Repairing Goldman's Group Scoring Rule. *Episteme*, *9*(2), 101–114. http://doi.org/10.1017/epi.2012.3

Kopec, M. (2017). A pluralistic account of epistemic rationality. *Synthese*, (March), 1–30. http://doi.org/10.1007/s11229-017-1388-x

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118.

Kornblith, H. (1999). Distrusting Reason. *Midwest Studies in Philosophy*, *23*(1), 181–196. http://doi.org/10.1111/1475-4975.00010

Kraft, P. W., Lodge, M., & Taber, C. S. (2015). Why People "Don't Trust the Evidence": Motivated Reasoning and Scientific Beliefs. *Annals of the American Academy of Political and Social Science*, *658*(1), 121–133. http://doi.org/10.1177/0002716214554758

Krause, S., James, R., Faria, J. J., Ruxton, G. D., & Krause, J. (2011). Swarm intelligence in humans: Diversity can trump ability. *Animal Behaviour*, *81*(5), 941–948. http://doi.org/10.1016/j.anbehav.2010.12.018

Krebs, J., & Dawkins, R. (1984). Animal signals: mind-reading and manipulation. *Behavioural Ecology: An Evolutionary Perspective*.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing". *Psychological Review*, *103*(2), 263–83.

Kuhn, D., & Lao, J. (1996). Effects of Evidence on Attitudes: Is Polarization the Norm? *Psychological Science*, *7*(2), 115–120. http://doi.org/10.1111/j.1467-9280.1996.tb00340.x

Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of Dyadic Interaction on Argumentative Reasoning. *Cognition and Instruction*, *15*(3), 287–315. http://doi.org/10.1207/s1532690xci1503_1

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, *53*(4), 636–647. http://doi.org/10.1037/0022-3514.53.4.636

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. http://doi.org/10.1037/0033-2909.108.3.480

Lackey, J. (2005). Memory as a Generative Epistemic Source. *Philosophy and Phenomenological Research*, *70*(3), 636–658.

Lackey, J. (2008a). *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.

Lackey, J. (2008b). A Justificationist View of Disagreement's Epistemic Significance. In A. Haddock, A. Millae, & D. Pritchard (Eds.), *Social Epistemology*. Oxford: Oxford University Press. http://doi.org/10.1093/acprof

Lackey, J. (2010). What Should We Do When We Disagree? *Oxford Studies in Epistemology 3*, 274–293.

Lam, B. (2011). On the Rationality of Belief-Invariance in Light of Peer Disagreement. *Philosophical Review*, *120*(Feldman 2007), 207–245. http://doi.org/10.1215/00318108-2010-028

Lam, B. (2013). Calibrated probabilities and the epistemology of disagreement. *Synthese*, *190*(6), 1079–1098. http://doi.org/10.1007/s11229-011-9881-0

Landemore, H. (2012). Deliberation, cognitive diversity, and democratic inclusiveness: an epistemic argument for the random selection of representatives. *Synthese*, *190*(7), 1209–1231. http://doi.org/10.1007/s11229-012-0062-6

Lasonen-Aarnio, M. (2014). Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, *88*(2), 314–345. http://doi.org/10.1111/phpr.12090

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and Social Combination Processes on Mathematical Intellective Tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189.

Lessig, L. (1995). The Regulation of Social Meaning. *The University of Chicago Law Review*, *62*(3), 943–1045.

Levy, N. (2006). Open-Mindedness and the Duty to Gather Evidence. *Public Affairs Quarterly*, *20*(1), 55–66.

Lewandowsky, S., Gignac, G. E., & Oberauer, K. (2013). The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS ONE*, *8*(10). http://doi.org/10.1371/journal.pone.0075637

Liberman, A., & Chaiken, S. (1992). Defensive Processing of Personally Relevant Health Messages. *Personality and Social Psychology Bulletin*, *18*(6), 669–679. http://doi.org/0803973233

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology Human Learning and Memory*, *4*(6), 551–578. http://doi.org/10.1037/0278-7393.4.6.551

Lindeman, M. (2011). Biases in intuitive reasoning and belief in complementary and alternative medicine. *Psychology and Health*, *26*(3), 371–382. http://doi.org/10.1080/08870440903440707

Littlejohn, C. (2012). *Justification and the Truth Connection*. Cambridge University Press.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. http://doi.org/10.1037/0022-3514.47.6.1231

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

Luskin, R. C., O'Flynn, I., Fishkin, J. S., & Russell, D. (2012). Deliberating across Deep Divides. *Political Studies*, *62*(1), 116–135. http://doi.org/10.1111/j.1467-9248.2012.01005.x

Mann, R. P., & Helbing, D. (2016). Minorities report: optimal incentives for collective intelligence. *Proceedings of the National Academy of Sciences*, *114*(20), 5077–5082. http://doi.org/10.1073/pnas.1618722114

Manza, J., & Cook, F. L. (2002). A Democratic Polity? Three Views of Policy Responsiveness to Public Opinion in the United States. *American Politics Research*, *30*(6), 630–667. http://doi.org/10.1177/153267302237231

Matheson, J. D. (2009). Conciliatory Views of Disagreement and Higher-Order Evidence. *Episteme*, *6*(3), 269–279. http://doi.org/10.3366/E1742360009000707

Matheson, J. D. (2014). A Puzzle About Disagreement and Rationality. *Social Epistemology Review and Reply Collective*, *3*(4), 1–3.

Matheson, J. D. (2015). *The epistemic significance of disagreement*. Palgrave Innovations in Philosophy. http://doi.org/10.1007/s13398-014-0173-7.2

Mayo-Wilson, C., Zollman, K., & Danks, D. (2013). Wisdom of crowds versus groupthink: Learning in groups and in isolation. *International Journal of Game Theory*, *42*(3), 695–723. http://doi.org/10.1007/s00182-012-0329-7

McCright, A. M., & Dunlap, R. E. (2011). The Politization if Climate Change and Polarization in the American Public's Views of Global Warming, 2001-2010. *The Sociological Quarterly*, *52*(2), 155–194.

McCright, A. M., Xiao, C., & Dunlap, R. E. (2014). Political polarization on support for government spending on environmental protection in the USA, 1974-2012. *Social Science Research*, *48*(November), 251–260. http://doi.org/10.1016/j.ssresearch.2014.06.008

McGarity, T. O. (2002). Professor Sunstein's Fuzzy Math. *The Georgetown Law Journal*, *90*, 2341–2377.

McPherson Frantz, C., & Janoff-Bulman, R. (2000). Considering Both Sides: The Limits of Perspective Taking. *Basic and Applied Social Psychology*, *22*(1), 31–42. http://doi.org/10.1207/S15324834BASP2201_4

McQueen, P. (2017). When Should we Regret? *International Journal of Philosophical Studies*, *25*(5), 608–623. http://doi.org/10.1080/09672559.2017.1381408

Mercier, H. (2011). What good is moral reasoning? *Mind Society*, *10*(2), 131–148. http://doi.org/10.1007/s11299-011-0085-6

Mercier, H., Deguchi, M., Van der Henst, J.-B., & Yama, H. (2015). The benefits of argumentation are cross-culturally robust: The case of Japan. *Thinking & Reasoning*, (January), 1–15. http://doi.org/10.1080/13546783.2014.1002534

Mercier, H., & Landemore, H. (2012). Reasoning Is for Arguing: Understanding the Successes and Failures of Deliberation. *Political Psychology*, *33*(2), 243–258. http://doi.org/10.1111/j.1467-9221.2012.00873.x

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, *34*(2), 57-74-111. http://doi.org/10.1017/S0140525X10000968

Mercier, H., Trouche, E., Yama, H., Heintz, C., & Girotto, V. (2015). Experts and laymen grossly underestimate the benefits of argumentation for reasoning. *Thinking & Reasoning*, *21*(3), 341–355. http://doi.org/10.1080/13546783.2014.981582

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology*, *74*(5), 834–839.

Miller, A. G., McHoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change. *Journal of Personality and Social Psychology*, *64*(4), 561–574.

Moffett, M. (2007). Reasonable Disagreement and Rational Group Inquiry. *Episteme*, *4*(January 2012), 352–367. http://doi.org/10.3366/E1742360007000135

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, *4*(3), 231–248.

Moss, S. (2015). Time–Slice Epistemology and Action under Indeterminacy. *Oxford Studies in Epistemology, Volume 5*, *5*, 172–193. http://doi.org/10.1093/acprof

Muldoon, R. (2013). Diversity and the Division of Cognitive Labor. *Philosophy Compass*, *8*(2), 117–125. http://doi.org/10.1111/phc3.12000

Munro, G. D., & Ditto, P. H. (1997). Biased Assimilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information. *Personality and Social Psychology Bulletin*, *23*(6), 636–653.

Munro, G. D., Ditto, P. H., Lockhart, L. K., Fagerlin, A., & Gready, M. (2002). Biased Assimilation of Sociopolitical Arguments: Evaluating the 1996 U.S. Presidential Debate. *Basic and Applied Social Psychology*, *24*(1), 15–26. http://doi.org/10.1207/S15324834BASP2401

Munro, G. D., & Stansbury, J. a. (2009). The dark side of self-affirmation: confirmation bias and illusory correlation in response to threatening information. *Personality and Social Psychology Bulletin*, *35*(9), 1143–1153. http://doi.org/10.1177/0146167209337163

National Research Council. (2012). *Deterrence and the Death Penalty*. Washington, D.C.: National Academies Press. http://doi.org/10.17226/13363

Nel, L. D., & Steele, C. M. (2000). Do Messages About Health Risks Threaten the Self ? Increasing the Acceptance of Threatening Health Messages Via Self-Affirmation, *26*(9), 1046–1058. http://doi.org/10.1177/01461672002611003

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. http://doi.org/10.1037/1089-2680.2.2.175

Nisbett, R., & Wilson, T. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259.

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political

Misperceptions. *Political Behavior*, *32*(2), 303–330. http://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, *133*(4), e835-42. http://doi.org/10.1542/peds.2013-2365

Olsson, E. (2017). Polarization - Is it rational? In *Workshop on Groups and Disagreement, University of Copenhagen*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. http://doi.org/10.1126/science.aac4716

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: how a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press.

Park, H. S., Levine, T. R., Westerman, C. Y. K., Orfgen, T., & Foregger, S. (2007). The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social judgment predictions. *Human Communication Research*, *33*(1), 81–102. http://doi.org/10.1111/j.1468-2958.2007.00290.x

Paxton, J. M., Ungar, L., & Greene, J. D. (2011). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163–77. http://doi.org/10.1111/j.1551-6709.2011.01210.x

Peters, E., & Slovic, P. (1996). The Role of Affect and Worldviews as Orienting Dispositions in the Perception and Acceptance of Nuclear Power. *Journal of Applied Social Psychology*, *26*(16), 1427–1453. http://doi.org/10.1111/j.1559-1816.1996.tb00079.x

Petersen, M. B., Skov, M., Serritzlew, S., & Ramsøy, T. (2012). Motivated Reasoning and Political Parties: Evidence for Increased Processing in the Face of Party Cues. *Political Behavior*, *35*(4), 831–854. http://doi.org/10.1007/s11109-012-9213-1

Pettit, P. (2006). When to defer to majority testimony - and when not. *Analysis*, *66*(291), 179–187. http://doi.org/10.1111/j.1467-8284.2006.00612.x

Petty, R. E., Wegener, D. T., & Fabrigar, L. R. (1997). Attitudes and attitude change. *Annual Review of Psychology*, *48*, 609–47. http://doi.org/10.1146/annurev.psych.48.1.609

Pew Research Center. (2015). Public and Scientists' Views on Science and Society, *January 29*.

Pew Research Center. (2016). *The Politics of Climate* (Vol. October 4). Retrieved from http://jwelb.oxfordjournals.org/cgi/doi/10.1093/jwelb/jwp029

Plantinga, A. (2000). *Warranted Christian Belief*. Oxford University Press.

Plous, S. (1991). Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer? *Journal of Applied Social Psychology*, *21*(13), 1058–1082.

Polzer, J. T., Milton, L. P., & Swann, W. B. (2002). Capitalizing on Diversity: Interpersonal

Congruence in Small Work Groups. *Administrative Science Quarterly*, *47*(2), 296. http://doi.org/10.2307/3094807

Pomerantz, E. M., Chaiken, S., Tordesillas, R. S., Chen, S., Darke, P., Eagly, A., … Zimmerman, J. (1995). Attitude Strength and Resistance Processes. *Journal of Personality and Social Psychology*, *69*(3), 408–419.

Price, M. E. (2012). Group Selection Theories are Now More Sophisticated , but are They More Predictive? *Evolutionary Psychology*, *10*(1), 45–49.

Price, V., Cappella, J. N., & Nir, L. (2002). Does Disagreement Contribute to More Deliberative Opinion? *Political Communication*. http://doi.org/10.1080/105846002317246506

Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press. http://doi.org/10.1093/019928038X.001.0001

Pronin, E., Lin, D. Y., & Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. http://doi.org/10.1177/0146167202286008

Quine, W. V. O. (1969). Epistemology naturalized. In W. V. O. Quine (Ed.), *Ontological Relativity and Other Essays* (pp. 69–90). New York: Columbia University Press.

Quong, J. (2018). Public Reason. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy* (Spring 201). Stanford University. Retrieved from https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=public-reason

Rawls, J. (1971). *A theory of justice*. Cambridge, Ma.: Belknap Press.

Rawls, J. (1993). *Political liberalism*. New York, NY: Columbia University Press.

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-process approach. *Current Directions in Psychological Science*, *13*(2), 60–66.

Richey, M. (2012). Motivated Reasoning in Political Information Processing: The Death Knell of Deliberative Democracy? *Philosophy of the Social Sciences*, *42*(4), 511–542. http://doi.org/10.1177/0048393111430761

Rieznik, A., Moscovich, L., Frieiro, A., Figini, J., Catalano, R., Garrido, J. M., … Gonzalez, P. A. (2017). A massive experiment on choice blindness in political decisions : Confidence , confabulation , and unconscious detection of self-deception. *PloS One*, 1–16. http://doi.org/10.1371/journal.pone.0171108

Roberts, L. (1990). Counting on science at EPA. *Science*, *249*(4969), 616–8. http://doi.org/10.1126/SCIENCE.2382138

Roeser, S. (2010). *Emotions and risky technologies*. (S. Roeser, Ed.). Dordrecht: Springer.

Rosen, G. (2001). Nominalism, Naturalism, and Epistemic Relativism. *Philosophical*

Perspectives, *15*, 69–91.

Ross, L. (2012). Reflections on Biased Assimilation and Belief Polarization. *Critical Review*, *24*(2), 233–245. http://doi.org/10.1080/08913811.2012.711025

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. *Judgment under Uncertainty: Heuristics and Biases*, (1977), 129–153.

Santaella-Tenorio, J., Cerdá, M., Villaveces, A., & Galea, S. (2016). What Do We Know about the Association between Firearm Legislation and Firearm-Related Injuries? *Epidemiologic Reviews*, *38*(1), 140–157. http://doi.org/10.1093/epirev/mxv012

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology*, *91*(6), 1080–93. http://doi.org/10.1037/0022-3514.91.6.1080

Schulz-Hardt, S., Jochims, M., & Frey, D. (2002). Productive conflict in group decision making: Genuine and contrived dissent as strategies to counteract biased information seeking. *Organizational Behavior and Human Decision Processes*, *88*, 563–586. http://doi.org/10.1016/S0749-5978(02)00001-8

Schwartzman, M. (2011). The Sincerity of Public Reason. *Journal of Political Philosophy*, *19*(4), 375–398. http://doi.org/10.1111/j.1467-9760.2010.00363.x

Shafir, E., & Leboeuf, R. a. (2002). Rationality. *Annual Review of Psychology*, *53*, 491–517.

Sherman, D. K., & Cohen, G. L. (2002). Accepting Threatening Information: Self-Affirmation and the Reduction of Defensive Biases. *Current Directions in Psychological Science*, *11*(4), 119–123. http://doi.org/10.1111/1467-8721.00182

Sherman, D. K., Kinias, Z., Major, B., Kim, H. S., & Prenovost, M. (2007). The group as a resource: reducing biased attributions for group success and failure via group affirmation. *Personality and Social Psychology Bulletin*, *33*(8), 1100–1112. http://doi.org/10.1177/0146167207303027

Shi, J., Visschers, V. H. M., & Siegrist, M. (2015). Public Perception of Climate Change: The Importance of Knowledge and Cultural Worldviews. *Risk Analysis*, *35*(12), 2183–2201. http://doi.org/10.1111/risa.12406

Shih, T. J., Scheufele, D. A., & Brossard, D. (2013). Disagreement and value predispositions: Understanding public opinion about stem cell research. *International Journal of Public Opinion Research*, *25*(3), 357–367. http://doi.org/10.1093/ijpor/eds029

Sliwa, P., & Horowitz, S. (2015). Respecting all the evidence. *Philosophical Studies*, *172*(11),

2835–2858. http://doi.org/10.1007/s11098-015-0446-9

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings. *Risk Analysis*, *24*(2). Retrieved from papers2://publication/uuid/F4B3EA59-C2F7-4F62-9223-701C27F152C2

Smart, P. R. (2017). Mandevillian intelligence. *Synthese*, (April), 1–32. http://doi.org/10.1007/s11229-017-1414-z

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. http://doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2003). Evolutionary versus instrumental goals: How evolutionary psychology misconceives human rationality. *Evolution and the Psychology of Thinking: The Debate*, 171–223. http://doi.org/10.4324/9780203641606

Stanovich, K. E., & West, R. F. (2014). The Assessment of Rational Thinking: IQ != RQ. *Teaching of Psychology*, *41*, 265–271. http://doi.org/10.1177/0098628314537988

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478. http://doi.org/10.1037/0022-3514.48.6.1467

Strickland, A. A., Taber, C. S., & Lodge, M. (2011). Motivated Reasoning and Public Opinion. *Journal of Health Politics, Policy and Law*, *36*(6), 935–944. http://doi.org/10.1215/03616878-

Sunstein, C. R. (2002a). *Risk and reason: safety, law, and the environment*. Cambridge University Press.

Sunstein, C. R. (2002b). The Law of Group Polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

Sunstein, C. R. (2005). *Laws of Fear*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511790850

Sunstein, C. R. (2006). Misfearing: a reply. *Harvard Law Review*, *119*(4), 1110–1125.

Sunstein, C. R. (2014). *Valuing life: humanizing the regulatory state*. Chicago: The University of Chicago Press.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137–155. http://doi.org/10.1007/s11109-008-9075-8

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769.

Talbot, B. (2014). Truth promoting non-evidential reasons for belief. *Philosophical Studies*,

168(3), 599–618. http://doi.org/10.1007/s11098-013-0139-1

Thompson, V., & Evans, J. S. B. T. (2012). Belief bias in informal reasoning. *Thinking & Reasoning*, *18*(3), 278–310. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/13546783.2012.670752

Titelbaum, M. G. (2013). Rationality's fixed point (or: In defense of right reason). *Oxford Studies in Epistemology*, *5*, 253. http://doi.org/10.1093/acprof

Titelbaum, M. G., & Kopec, M. (2017). When Rational Reasoners Reason Differently. *Reasoning: Essays on Theoretical and Practical Thinking*, 1–25.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. http://doi.org/10.1080/13546783.2013.844729

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2016). The Selective Laziness of Reasoning. *Cognitive Science*, *40*(8), 2122–2136. http://doi.org/10.1111/cogs.12303

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. http://doi.org/10.1037/a0037099

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. http://doi.org/10.1126/science.7455683

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief The Role of Identity in Political Belief. *Trends in Cognitive Sciences*, *xx*, 1–12. http://doi.org/10.1016/j.tics.2018.01.004

van Fraassen, B. C. (1984). Belief and the Will. *Journal of Philosophy*, *81*(5), 235–256.

van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, *58*, 515–541. http://doi.org/10.1146/annurev.psych.58.110405.085546

Vinokur, A., & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology*, *36*(8), 872–885. http://doi.org/10.1037/0022-3514.36.8.872

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(February 2015), 273–281. http://doi.org/10.1080/14640746808400161

Wasserman, E. a, Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *16*(3), 509–521. http://doi.org/10.1037/0278-7393.16.3.509

Watson, W. E., Kamalesh, K., & Michaelsen, L. K. (2016). Cultural Diversity's Impact on Interaction Process and Performance : Comparing Homogeneous and Diverse Task Groups. *Academy of Management Journal*, *3*(3), 590–602.

Weatherson, B. (2016). *Normative Externalism*. Unpublished manuscript.

Weber, M. (1964). *The Theory Of Social And Economic Organization*. (T. Parsons, Ed.). New York: Free Press.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199251315.001.0001

White, R. (2005). Epistemic Permissiveness. *Philosophical Perspectives*, *19*(1), 445–459. http://doi.org/10.1111/j.1520-8583.2005.00069.x

White, R. (2010). You Just Believe That Because…. *Philosophical Perspectives*, *24*(1), 573–615. http://doi.org/10.1111/j.1520-8583.2010.00204.x

Wilson, T. D., Kraft, D., & Dunn, D. S. (1989). The disruptive effects of explaining attitudes: The moderating effect of knowledge about the attitude object. *Journal of Experimental Social Psychology*, *25*(5), 379–400. http://doi.org/10.1016/0022-1031(89)90029-2

Wojcieszak, M., & Price, V. (2010). Bridging the Divide or Intensifying the Conflict? How Disagreement Affects Strong Predilections about Sexual Minorities. *Political Psychology*, *31*(3), 315–39. http://doi.org/10.1111/j.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. http://doi.org/10.1126/science.1193147

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective intelligence and group performance. *Current Directions in Psychological Science*, *24*(6), 420–424. http://doi.org/10.1177/0963721415599543

Aasen, M. (2017). The polarization of public concern about climate change in Norway. *Climate Policy*, *17*(2), 213–230. http://doi.org/10.1080/14693062.2015.1094727